Final Technical Report for DE-FG06-92ER61487
Maynard Olson, Principal Investigator

I. INTRODUCTION

This project has been oriented toward improving the vertical integration of the sequential steps associated with the large-scale analysis of human DNA. The central focus has been on an approach to the preparation of "sequence-ready" maps, which is referred to as multiple-complete-digest (MCD) mapping, primarily directed at cosmid clones. MCD mapping relies on simple experimental steps, supported by advanced image-analysis and map-assembly software, to produce extremely accurate restriction-site and clone-overlap maps. We believe that MCD mapping is one of the few high-resolution mapping systems that has the potential for high-level automation. Successful automation of this process would be a landmark event in genome analysis. Once automated, MCD mapping could be applied widely to the human genome and to the genomes of other higher organisms, paving the way for cost-effective sequencing of these genomes. Critically, MCD mapping has the potential to provide built-in quality control for sequencing accuracy and to make possible a highly integrated end product even if there are large numbers of discontinuities in the actual sequence.

High-resolution physical mapping of complex genomes poses continuing challenges. Early successes in mapping the genomes of E. coli (Kohara et al., 1987), S. cerevisiae (Olson et al., 1986; Riles et al., 1993), and C. elegans (Coulson et al., 1988, 1991) led to maps of sufficient quality to support large-scale sequencing of the genomes of these organisms. Similar efforts directed at human chromosomes (Carrano et al., 1989; Stallings et al., 1990, 1992) have also enjoyed considerable success. Nonetheless, essentially complete, well organized sets of cosmids spanning tens of megabase pairs of human DNA have been slow to emerge. The pace at which large-scale-sequencing capacity is expanding threatens to overwhelm the genome-analysis community's collective ability to supply well mapped cosmid clones. This phenomenon is all the more overwhelming because of current interest in "low-pass" sequencing strategies since these strategies depend on particularly precise mapping if they are to avoid an endpoint so chaotic as to undermine the very rationale for gene discovery through the sequencing of germline DNA.

In retrospect, there are several reasons why DNA sequencing has advanced more rapidly than high-resolution physical mapping. Perhaps the most important cause has simply been that sequencing technology has received more attention than mapping technology. As discussed in the introduction, there is a substantial commercial market for reagents, instrumentation, and software to support DNA sequencing. Most of the customers who drive this market are sequencing short segments of DNA and require little commercial support for physical mapping. Hence, there is no corresponding market for large-scale mapping reagents, instrumentation, and software. The development of these research tools has largely been left to public-sector research laboratories. Furthermore, the laboratories with the greatest mapping expertise have been absorbed with the immense challenges posed by the low-resolution mapping of the human genome.

Some perspective on the technical issues underlying large-scale physical mapping can be gained by examining the relationship between the basic steps required to map a genome with those required

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

to sequence a cosmid by a shotgun method. Viewed superficially, two activities are quite similar. In both cases, the root steps are library construction, isolation of DNA from randomly picked clones, enzyme reactions, electrophoresis, image analysis, and computer-based assembly of a composite map or sequence. However, there are technical issues at nearly every step that pose special problems for high-resolution mapping, problems that are often more difficult than their sequencing counterparts. A step-by-step development of this perspective is provided below:

i. Library construction. Cloning of cosmid-sized segments of exogenous DNA puts more stress on the compatibility between the cloned DNA and the replication and recombination machinery of E. coli than does the cloning of plasmid-sized segments. Undoubtedly, limitations on cosmid-contig sizes are partly due to problems with the viability or stability of clones derived from specific regions (Coulson et al., 1991). Surprisingly little is known about the bacterial genetics--or even the physiology--of clone instability. Although many specialized hosts and vectors have been developed with the goal of improving the viability and stability of troublesome clones, judgments about the effectiveness of these measures rest almost entirely on anecdotal evidence.

ii. DNA isolation. It is more difficult to prepare pure DNA reliably from cosmid-sized clones than from small plasmids. Cosmid chromosomes have a cross section for suffering breaks or nicks that is an order of magnitude larger than that of typical plasmids. Broken or nicked molecules are nearly impossible to separate from E. coli chromosomal DNA since the best separation methods depend on maintaining the cosmid DNA in supercoiled form. Cosmid-sized DNA molecules also show more irreversible binding to solid absorbents and more entrapment in denatured biomass than do plasmid-sized molecules. Once again, a major obstacle to progress is the lack of a solid literature documenting the tradeoffs inherent in different approaches to cosmid-DNA purification. In relation to sequencing, it should be noted that double-stranded DNA is already more difficult to prepare in pure form even from small plasmids than is the viral DNA of filamentous-phage clones. Finally, PCR offers an alternative route to microgram amounts of the inserts of plasmid-sized clones, while cosmid DNA must still be amplified in vivo.

iii. Enzyme reactions. As a byproduct of PCR technology, linear amplification of sequencing products ("cycle sequencing") has lowered the amounts of template DNA required for DNA sequencing by an order of magnitude. There is no comparable strategy for lowering the amount of cosmid DNA required for standard fingerprinting procedures.

iv. Electrophoresis. Typical mapping methods can employ electrophoretic techniques that are as simple as, or simpler than, those used for DNA sequencing. However, DNA sequencing typically requires fractionation of DNA molecules that vary in size by a factor of little more than 10. In contrast, it is not unusual, when fractionating restriction digests, to need to separate molecules varying in size by a factor of 100. This need for a large dynamic range in size resolution stresses the capability of any single electrophoretic system.

v. Image analysis. The great success of "base-calling" methods in four-color-fluorescence sequencing depends on the existence of an orderly "ladder" of bands. Mapping methods based on complete digestion yield fragments whose sizes are uncorrelated with one another and whose size distribution is typically geometric, leading to clusters of small fragments. Within these clusters, the image-analysis methods must reliably identify co-migrating fragments.

vi. Shotgun assembly. Sequence-assembly procedures determine clone overlaps on the basis of much more information than do map-assembly procedures. In particular, complete-digest-assembly methods must determine the order of the fragments at the same time that they determine the order of the clones.

Taken individually, there are effective solutions to most of the problems cited above. For example, the need for pure DNA can be circumvented by direct- or indirect-end-labeling methods that combine extreme sensitivity with various sources of specificity (Coulson et al., 1988; Kohara et al., 1987; Carrano et al., 1989). A variety of methods, such as repetitive-sequence hybridization (Stallings et al. 1990) can be used to supplement fragment-size data with additional information about the fragments in a fingerprint. The size range over which good electrophoretic resolution can be achieved can be extended by methods such as field-inversion gel electrophoresis (Carle et al., 1986; Graham et al., 1987; Birren et al., 1989; Lai et al., 1991). In some mapping systems, instrumentation designed for DNA sequencing can be used to acquire the needed gel images (Carrano et al., 1989; Lamerdin and Carrano, 1993), thereby harvesting the large commercial investment in sequencing technology for mapping purposes. Partial-digestion methods can be used to simplify map assembly by allowing direct determination of fragment order, independently of clone order (Smith and Birnstiel 1976, Rackwitz et al., 1984; Kohara et al., 1987; Takahashi-Fujii et al., 1994). Finally, fingerprinting of clones with restriction enzymes can be bypassed altogether. For example, mapping techniques that depend on clone-overlap detection on the basis of DNA-DNA hybridization obviate the need for enzyme reactions and electrophoresis (Craig et al., 1990; Mizukami et al., 1993).

In summary, cosmid-level mapping poses a set of distinctive and challenging technical problems. A rich repertoire of potential solutions to these problems has been devised. However, these solutions involve a complex web of tradeoffs. Mixing and matching solutions to component problems is of limited value since what is needed is a complete, integrated strategy whose parts work well together. Even after more than a decade of work, no dominant technology has emerged. While the lack of a dominant technology favors innovation, it does not favor technological refinement. The contrast with DNA sequencing is striking: dideoxy sequencing became strongly dominant 15 years ago and the 4-color-fluorescence implementation had captured most of the high-end market by 1990. Hence, tremendous energy has gone into the refinement of these dominant technologies in recent years.

For this project, we have adopted multiple-complete-digest (MCD) mapping as our basic approach. An extensive discussion of the relative merits of this strategy and various alternatives is provided in the Introduction to our manuscript (Gillett et al., 1996). MCD mapping is a direct extension of the single-complete-digest method employed to map the S. cerevisiae genome (Olson et al., 1986; Riles et al., 1993). Major strengths and weaknesses of MCD mapping are summarized below:

Strengths

---Experimental simplicity: data collection involves restriction digestion, agarose gel electrophoresis, and bulk-staining of gels with fluorescent dyes.

---Reliance on model-based data acquisition and analysis: digests produce fragments in equimolar amounts whose sizes must sum to the size of the clone; maps must be unbranched and account for all fragments.

---Accuracy of maps: gross errors are rare because of the extensive internal cross-checks available when a genome is sampled at high redundancy; site coordinates are also established with high accuracy.

---Extensibility to arbitrarily complex genomes: fingerprints with two or three independent enzymes contain sufficient information to define unique clone placements in the largest known genomes unless large-scale, nearly exact genome duplications are present.

---Scalability of map resolution: enzymes with different levels of specificity can be matched to clones of different sizes to support mapping with resolutions varying from 0.1-100 kbp with no basic change in instrumentation, protocols, or software; at low resolution, this feature offers the prospect for long-range continuity in the maps, overlapping the resolution achievable by methods such as YAC-based STS-content mapping; at high-resolution MCD mapping offers a path toward directed sequencing.

Weaknesses

---Reliance on pure DNA: the lack of specificity associated with bulk staining requires significant quantities of pure DNA.

---Dependence on high-quality data: the intolerance of inconsistencies between maps and the underlying digest data prevents utilization of sub-standard data.

---Complexity of map-assembly problem: the need to order fragments and clones simultaneously, to maintain synchrony across different digest domains, and to impose absolute topology checking demands extremely sophisticated software.

Collectively, these strengths and weaknesses describe a method that is poorly suited for small-scale applications but that is a good candidate for high-level automation. We believe that MCD mapping is solidly in the running to become a dominant high-resolution mapping method, one that could drive the sequencing of mammalian and other complex genomes. The method is not competitive with any number of alternatives for small or modest-sized problems. However, its experimental simplicity, reliance on model-based data-acquisition and analysis, and high map quality are hallmarks of a method that is worth a major development effort, particularly given the staggering demand for efficient, high-resolution, cosmid-level mapping that lies at the core of the Human Genome Project.

II. PROGRESS ACHIEVED UNDER DE-FG06-92ER61487

MCD mapping has been reduced to practice for problems of modest size. Most of our experience involves subcloning YACs into cosmids and fingerprinting the cosmids with three enzymes that recognize 6-bp recognition sites. The digests are separated on agarose gels, which are stained with ethidium bromide, thiazole orange (Lee et al., 1986), or SYBR green (Molecular Probes, Eugene, OR). Images are acquired on a FluorImager 575 (Molecular Dynamics, Sunnyvale, CA) and analyzed using custom-developed image-analysis software. Map assembly is carried out by a custom-developed software package developed in collaboration with W. Gillett of the Washington University Department of Computer Science (Gillett et al. 1996).

The status of individual components of the MCD-mapping system is reviewed below. The bottom line is that the system works and produces truly excellent maps. However, significant development is still required to improve data quality, decrease the labor-intensiveness of the experimental steps, and to decrease the need for expert intervention during data analysis. Of these problems, the last is certainly the most daunting. There are many parallels with DNA sequencing, in which the "finishing" costs of shotgun-sequencing projects often account for a large fraction of the total cost. "Finishing" is simply a euphemism for highly customized expert intervention to compensate for the failure of standardized procedures to yield adequate results. As discussed in the sequencing project, major progress has been made in decreasing the need for custom finishing during shotgun sequencing. Similar comments apply to MCD mapping except that the system is less fully developed than shotgun sequencing.

A. COSMID-LEVEL MCD MAPPING

1. Upstream Steps

Our first substantial application of MCD mapping to human DNA has involved subcloning YACs from the Class I HLA region and carrying out MCD mapping on the resultant cosmids. This project, in collaboration with D. Geraghty (Fred Hutchinson Cancer Research Center), takes advantage of Geraghty's excellent YAC-based STS-content map of the class I region (Geraghty et al., 1992; Geraghty, 1993). Subcloning YACs into cosmids has the disadvantage that aberrations in the YACs could be passed all the way to the sequence level without detection. However, this approach has the offsetting advantage that it allows a ready source of deep cosmid coverage without requiring an up-front commitment to particular cloning technology. Indeed, most published high-resolution maps of YAC-sized regions of the human genome have been prepared from cosmid or lambda clones subcloned from YACs (Whittaker et al., 1993; Pieretti et al., 1991). During the HLA Class I project, we plan to protect ourselves against the propagation of YAC artifacts by a combination of redundant mapping from independent YACs and the use of RARE cleavage to compare the contour lengths of our maps with those present in a redundant set of YACs. Neither of these methods is practical for projects appreciably larger than the HLA Class I region (2+ Mbp); however, they are adequate to support our immediate need for an approach that allows us to interact directly with large-scale sequencing projects while still conducting extensive experimentation with cosmid vectors, hosts, and growth protocols.

Most of the protocols that we employ for cosmid cloning, hybridization-based screening of cosmid libraries, cosmid-DNA preparations, and cleavage of cosmids with restriction enzymes are standard. However, it has been necessary to modify standard cosmid vectors. This need arises because convenient MCD mapping places constraints on the restriction maps of the vectors. Our goal has been to develop a repertoire of enzymes with 6-bp recognition sites that are relatively inexpensive, give good practical results, and do not cleave within vector sequences. When there are no sites for the mapping enzymes within the vector, cosmid digests produce a single vector-containing fragment, which is fused to an indeterminate amount of DNA derived from both ends of the insert. At present, the identity of this fragment is determined by gel-transfer hybridization. However, now that we are meeting our goal of having a repertoire of enzymes suitable for mapping that produce a single vector-containing fragment we plan to develop software-level approaches to vector-band identification.

Our starting cosmid vector was sCos-1 (Evans et al., 1989). The neomycin-resistance gene and SV40 replication origin have been deleted, eliminating the only HindIII site in the vector. The resultant derivative, sCos-d, is suitable for cloning of MboI partial-digest fragments into a BamH1 site. HindIII, BglII, NsiI, and NdeI are all good mapping enzymes for which sCos-d lacks sites. A further derivative of sCos-d, sCos-DRI, contains a modified cloning site suitable for cloning of partial-digest fragments prepared with EcoRI or ApoI. In addition to the enzymes suitable for use with sCos-d, EcoRI is a good mapping enzyme for DNA cloned into sCos-DRI. Furthermore, this vector is ideal for implementation of RARE-cleavage-based contig anchoring (see below).

2. Electrophoresis

The basic conditions that we employ for electrophoresis have changed little in many years. The best results are obtained on thermostated 1.0-1.5% agarose gels run slightly below ambient temperature in a high-salt buffer that is recirculated vigorously (Helms et al., 1987). Some preliminary data have been collected with TBE as running buffer; this low-salt buffer leads to better separation between RNA and small double-stranded DNA fragments, an important consideration when the gels are stained with thiazole orange, a dye that stains RNA particularly intensely. However, it is clear that high-salt buffers give the best band-sharpness and lane-to-lane consistency even though their higher electrical conductance puts more stress on the temperature-control system.

The interplay between the performance of the DNA preparations, the electrophoretic instrumentation, the staining protocol, and the image-analysis software illustrates a general feature of this type of technology development. Optimization of individual steps, taken in relative isolation, is ineffective. This realization has driven us toward the integrated, team-oriented management system described in the Administrative Core section, despite its unorthodoxy in an academic environment.

To standardize electrophoretic conditions, particularly as we scale up mapping to a level where many gels must be run simultaneously, we have designed and built a standardized apparatus suitable for replication. This apparatus runs 3 stacked gels simultaneously, employs pressed-alumina gel plates and heat-exchange components, monitors the temperature with thermistors at many sites within the electrophoretic chamber, and is fully computer-controlled. The heavy use of pressed-alumina ceramic components provides the best available compromise between high heat conductivity and electrolytic inertness. By putting the entire control system in software, we expect to be able to introduce steady improvements in temperature and, perhaps, electric-field control without significant re-engineering of the actual apparatus. The redesigned apparatus is just coming into use: early indications are that it already performs as well or better than its jerry-rigged predecessor, which was passed down from the yeast-mapping project.

3. Gel Staining

The gel-staining protocol has emerged as a critical experimental step with implications for many aspects of the overall process. For example, the sensitivity achieved by bulk-staining of the gels determines the physical scales of the bacterial-growth and DNA-extraction steps. We have experimented extensively with three intercalating dyes, ethidium bromide, thiazole orange, and SYBR green. When gels are scanned with the Molecular Dynamics FluorImager 575, thiazole orange and SYBR green both outperform ethidium bromide by a wide margin. Comments about the characteristics of these two preferred dyes are presented below.

Thiazole orange gives optimum results when digests are carried out on 100-ng samples of cosmid DNA for our typical lane cross sections (3 mm x 8 mm). Thiazole orange is cheap, relatively sensitive, and gives excellent linearity in double-log plots of fluorescent intensity vs. DNA mass over a wide mass range (see below, for comment about the empirical form of the intensity-mass relationship). The main disadvantage is relatively poor specificity for double-stranded DNA. Not only does RNA stain intensely, but thiazole-orange images are generally dirty due to intense staining of particulate impurities such as dust.

SYBR green has a number of advantages over thiazole orange. It has several-fold higher sensitivity and much greater specificity. Indeed, it provides visually stunning gel images because of the low background fluorescence. The optimum loading level is approximately 20 ng of cosmid DNA, although the exceptional signal-to-noise characteristics of this staining system allows usable cosmid fingerprints to be obtained with as little as 2 ng of cosmid DNA. One practical disadvantage is the relatively high cost of SYBR green ($10/gel, as opposed to $0.04/gel for thiazole orange); obviously, this high cost is related to SYBR green's status as a proprietary reagent. A technical disadvantage is that, for unknown reasons, intensity-mass plots for SYBR green begin to saturate when the mass of DNA per band is approximately 10 ng in our standard gel geometry, a threshold that is insensitive to staining conditions. Hence, good results are only obtained over a narrow range of DNA-loading levels, in contrast to thiazole orange. Because of widespread interest in this important new dye, we are assuming that more solid information will emerge about its chemical and physical properties. However, at present, even the molecular structure of SYBR green is a trade secret. We will be cautious about increasing our reliance on this dye until the situation changes

4. Image Analysis

Large-scale applications of MCD mapping will rely critically on automatic extraction of fragment sizes from digital images of gels. Human involvement in gel interpretation is not only slow and unreliable, but even modest levels of success depend on long experience and special skills. There is a considerable literature describing approaches to the computer-based interpretation of gels of the type employed in restriction-site mapping (Agard et al., 1981; Gray et al., 1984; Drury et al., 1990; Galat and Goldberg, 1987; Smith and Thomas, 1990; Fibeiro and Sutherland, 1991; Drury et al., 1992). However, none of these efforts has had much impact on actual mapping projects. Indeed, the only software that has had any significant practical utility has left most difficult decisions to user judgment.

During the past two years, G.K.S. Wong, working with M. Olson, has developed a fully automatic system for interpreting mapping gels that is now in daily use and has even been successfully ported to R. Waterston's sequencing center. Approximately 200 gel images consisting of 6000 gel lanes have now been processed by this system. Of course, the software still makes occasional errors, but it also makes many of the decisions correctly that human interpreters often get wrong. On balance, the quality of the results approaches that achievable by an expert human interpreter. The software is described here in some detail. Not only is it mission-critical in its own right, but it documents real experience with an approach to eliminating the need for expert judgment that underlies the Automation Project.

The strengths and weakness of the existing image-analysis package are as follows:

1) We assume that there are relatively few band in each lane and that these bands are randomly distributed. The program may fail if there are more than about 30 bands per lane, or if the same band pattern is repeated on every lane. Furthermore, the lanes must run reasonably straight.

2) Our software virtually never misses a band. Among all of the gel images processed so far, we know of only 1 or 2 bands that should have been called but were not. Faint bands that are often missed by human interpreters are routinely picked up by this software.

3) The number of fragments in a multiplet band is correctly called even when there are no neighboring singlets to compare it against, but the fragment counting software makes occasional errors when too many bands, generally about 4 or 5, are clustered together. We are confident that further improvements in the non-linear curvefitting procedures applied to band clusters can be introduced relatively simply.

4) Our software is tolerant of dirt on the gel image. Problems are restricted to the bottom of the gel where the bands have the lowest intensity. Once every 4 or 5 gels, an isolated piece of dirt will be identified as a digest band or, if it sits right on top of a band, cause the fragment count for that band to be too high. As this statistic is based on thiazole-orange staining, better results are expected once we switch to a dye like SYBR green which does not stain dirt as strongly. No amount of expert interpretation or computer analysis can save a data set when a big enough piece of dirt falls on top of a faint band. This muddy reality, rather than fundamental molecular properties of gels and DNA sets a lower bound on the practical size of mapping gels. Since band size scales with the gel size, while the size distribution of particulate noise is unchanged, and increasing fraction of the data would be obscured by dirt as gels were downsized.

5) Bad data is automatically rejected. Among the growing list of anomalies that we screen for are insufficient signal-to-noise ratios, chromosomal debris, incomplete digests, deleted clones, mixed clones, overloaded gels, and DNA degradation.

6) The most fragile piece of this software is the automatic marker-band identification, whose successful operation depends on too many ad hoc assumptions about particular sets of marker bands and is intolerant of significant changes in electrophoretic conditions. It should be possible to design size markers that allow us to achieve more robust results.

The analysis of an image is broken down into a series of distinct logical steps, with the most generic operations coming first At present, an image is carried through the 7 steps described below:

1) Delimit the region of interest by finding the gel borders and the loading wells. Find the lanes and then rotate the gel image so as to make the lanes as close to vertical as possible.

2) Produce a one-dimensional representation of each gel lane, using a median-averaging technique to reduce the effects of dirt. Identify all of the "significant" local maxima as potential DNA bands. It is inevitable that this procedure will produce artifacts, and subsequent steps must attempt to sift out the good peaks from the bad.

3) Assign sizes to the marker bands by doing a non-linear pattern match against the pattern from a reference gel. Interpolate between the marker lanes to assign sizes to the digest bands.

4) Fit each of the previously identified local maxima with a model peak function. Closely spaced bands are fit in a single operation as a band cluster; independent straight line backgrounds are used for each of these clusters to allow for arbitrary long-range background variations.

5) Resize the bands using the curve-fitted peak positions. Determine the number of fragments in each digest band by analyzing the trend in integrated peak intensity versus fragment size.

6) Decompose each multiplet band into its individual components. This procedure is still in development, but the intent is to resolve any problems in the preliminary fragment count by repeating the curve fits with much tighter constraints on such parameters as peak asymmetry than cold be applied in step 4 .

7) Make sure that the sum of the fragment lengths is consistent with expectations, not just on a clone-by-clone basis, but also across different digests of the same clone. Send data to mapping software.

Our software is implemented as a series of roughly 150 script files that are executed under MatLab, a commercial multi-platform numerical analysis development system. Most of the development work was done on a 486/DX2-66MHz PC running Windows/DOS, but this software has also been tested on two different variations of UNIX, an HP-9000 and a Sparc-10, with good results. Both ports were done across the Internet, and neither required more than a half hour of work. Since the MatLab binary data files generated at the end of each of the 7 steps described above are cross-platform compatible, we can even distribute the processing steps among different computers running different operating systems.

Two details of the analytical strategy that have significant implications for further development of the experimental system are described below:

i. Size-mobility calibration. At present, we rely on external size markers, typically run every 4-5 lanes. Size-mobility curves for individual marker lanes are developed by a least-squares cubicspline procedure that is a built-in MatLab feature. While, in principle, a different spline is needed for every lane, we have found that there is a simple relation between all of the calibration curves on a given gel: if S1 = f1(x) and S2 = f2(x) denote the fragment size S as a function f() of the DNA mobility x for two lanes run under similar electrophoretic conditions, the functions f1 and f2 are related by the transform, f1(x) = f2(a*x+b) to sub-pixel accuracy. In this formulation, a and b are lane-specific parameters. At present, values of a and b appropriate for a particular experimental lane are determined by segmental-linear interpolation between flanking size-marker lanes.

Our approach to automatic marker-band identification depends on the observation that the linear transform discussed above works surprisingly well even across gels. Thus, by recording the calibration spline for a representative gel, we can generate, to within a scale and a shift factor, the expected band pattern for any marker run under similar gel conditions. Hence, as long as two marker bands can be identified reliably by some marker-specific heuristic (e.g., the two most intense bands in specified regions of the band pattern), the lane-specific a and b parameters can be determined. Once they are determined, the expected pattern of the bands in the lane can be calculated quite precisely, providing a good basis for band-identification via a simple pattern match.

ii. Peak intensity vs. DNA mass. Within the experimental lanes, the software must be able to compute exactly how many complete-digest fragments are present in each of the bands called by the peak finder. The operative assumption is that there is an equimolar quantity of each restriction fragment in the complete digest. As long as the fluorescent intensity is a reproducible function of the DNA mass in a band, we should be able to use the trend in the integrated peak intensity vs. fragment size to compute the fragment count.

Our images are analyzed at a resolution of 5 pixels per mm, meaning that some of the bands at the top of the gel are represented by as few as 5 data points. To simply sum up these 5 data points would be too inaccurate an approach to peak integration. It is better to fit every band in the one-dimensional lane profile with a model-peak function and to integrate that function instead. The appropriate function is certainly not a Gaussian. In our hands, the best fits are obtained with a generalized Lorentzian of the form 1/(1+|z|^3), where z=(x-x0)/W{i}. Asymmetry is incorporated by using different width parameters W{t} and W{b} for the top and the bottom portions of each band.

We have found that the integrated intensities of singlet bands typically do not increase linearly with the fragment size. Under most experimental conditions, the larger fragments tend to be less intense than expected. Fortunately, to an excellent approximation, the intensities scale as the fragment size raised to some exponent E. Exponents in the range 0.7-1.0 are typical (note that E=1.0 implies strict proportionality between integrated peak intensity and DNA mass). The value of E for a particular gel is not highly reproducible even when staining conditions are standardized. However, the single-parameter-power-law dependence of integrated-peak intensity on size is extremely robust. This observation allows excellent interpretations of the multiplicity of bands in particular lanes even if these bands have no nearby neighbors as intensity-comparison standards.

The overall philosophy of the image-analysis package is to reject gel lanes that fail to conform to the expected model without attempting detailed analysis of what went wrong in a particular rejected lanes. This practice of checking whether or not data conform to expectations is the essence of model-based data analysis, a practice that we intend to use throughout the Automation Project. In the image-analysis package, we are presently able to reject essentially all bad gel lanes while sacrificing no more than 1-2% of useable ones.

5. Map Assembly

The appendicized preprint by Gillett et al. (1996) provides an extensive introduction to MCD mapping. The Olson-Gillett collaboration now reflects over 5 years of sustained effort to design and implement MCD-mapping software. Over 300,000 lines of ANSI C source code have been produced. The complexity of the MCD map-assembly problem arises from many sources. Some of the most consequential issues are itemized below:

i. Partitioning. Partitioning involves grouping clones into contigs and determining likely clone-order within the contigs without attempting absolute-fragment accounting. Indeed, with the exception of the single-digest software developed for the yeast-mapping project (Olson et al., 1986), most previous forays into complete-digest mapping have stopped after a level of analysis comparable to a single-digest analog of the multiple-digest partitioning step of the Gillett package. In MCD mapping, partitioning is a minor, but important, pre-processing step. As presently implemented, partitioning has approximately n**2 complexity, where n is the number of clones, since it involves exhaustive pairwise comparisons between the clones. Later steps can be carried out on individual partitions, whose sizes can be controlled by the stringency of the partitioning criteria. This divide-and-conquer strategy is essential since the computational complexity of later steps is exponential in the number of clones considered. However, since contig size does not grow with genome size--indeed, it undoubtedly decreases for a variety of practical reasons--overall compute time for steps after partitioning scales linearly with genome size.

ii. Clone-end compatibility. MCD mapping is based on the principle that the clone ends (more precisely, the ends of the cloned inserts) must have compatible positions in the maps produced in all digest domains. Specifically, MCD mapping imposes the requirement that there must exist an ordering of all the clone ends that is consistent with the partial-clone-end order implicit in each of the single-digest maps. There are a number of subtleties to this problem. For example, the case of clones prepared by partial digestion using an enzyme that is also employed in the mapping imposes more rigid constraints than the more usual case in which mapped sites cannot coincide with clone ends. A more subtle problem involves the effects of unregistered fragments (e.g., fragments too small to detect by the experimental procedure employed). Unregistered fragments in a particular digestion domain can cause apparent clone-end incompatibilities between maps when clone ends lie within these fragments.

iii. Fragment confusion. The central problem in MCD mapping is fragment confusion, the assumption that two fragments are the same when actually they are only similar in size. The most pernicious fragment-confusion problem is the "collapsed-fragment" case, which leads to maps that contain only one local instance of a fragment of a particular size class, when the correct map would contain two such fragments. Collapsed-fragment errors occur most commonly when two fragments of similar size occur approximately one clone-length apart. Elaborate fragment-splitting procedures have been developed that allow after-the-fact recognition and correction of collapsed-fragment errors. Hence, the MCD-mapping software is based on a greedy algorithm that fixes problems as they become apparent without backtracking.

iv. Stratified mapping. An intriguing feature of MCD mapping, which has been implemented but not yet tested in real applications, is the ability to eliminate temporarily any size class of fragments from the analysis of a particular partition. As long as the size class is eliminated consistently across the whole set of clones, which requires identification of natural discontinuities in the size distribution of the fragments, MCD mapping can proceed as though the fragments did not exist. Once an MCD map has been built, the excluded fragments can be re-introduced in a process referred to as "sprinkling." In simulated examples, many cases of fragment confusion can be avoided by employing stratified mapping. An intriguing possibility--albeit one that one pose complex implementation challenges--would be to employ stratified mapping to exclude temporarily a size class of fragments that is judged likely to contain image-analysis errors. In many instances, the expectation is that only one interpretation a lane image would give any solution during the fragment-sprinkling step.

v. Advanced simulations. Construction of the MCD-mapping package would not have been possible without a sophisticated simulation capability. The current simulator starts with a DNA sequence--real or simulated--and produces simulated data sets that include simulations of the statistically well behaved components of experimental error. Only by exercising the software on large numbers of simulated data sets has it been possible to debug this complex set of functions.

Viewed formally, MCD mapping can be regarded as an elaborate decoding problem, in which the goal is to reconstruct a message from a coded representation of it. Like many encryption problems, the coded representation of the message is easily calculated from the message itself (i.e., the role of the simulator), while the reverse calculation (i.e., the role of the mapping package) is difficult. At this level, the development of MCD-mapping software is essentially complete. Simulated data for a 1-Mbp genomic region, which contain realistic amounts of error in the fragment sizes, can often be mapped with the current software without any mistakes. These results depend on deep sampling of the region (typically 15X-20X coverage in cosmids is assumed). The mapping of a 1-Mbp region still requires a few hours of interactive time. User interaction is primarily needed to straighten out fragment-confusion problems that the software does not yet handle correctly and to deal with poor fragment-size values estimates associated with the tails of the assumed error distribution. If further development efforts were directed toward solving these problems, the need for interactive involvement in the mapping of simulated data could undoubtedly be nearly eliminated. However, the major obstacles to large-scale applications of this software to real data lie elsewhere, as discussed under the Research Plan section. Consequently, only minor effort is no devoted to cleaning up the last remnants of the formal decoding problem.

6. Result Validation

The Gillett et al. (1996) manuscript describes the most rigorous test of MCD mapping yet carried out on real data. By employing two probes spaced roughly one cosmid-length apart, DNA was sampled for a 115-kbp segment of yeast chromosome III from a deep, whole-genome cosmid library . A two-enzyme MCD map was constructed using EcoRI and HindIII as the mapping enzymes. The redundancy of the sampling was approximately 15X. Fragment sizing was sub-optimum both because the gels were run at low salt and because an early version of the image-analysis package was employed. All inconsistencies that prevented incorporation of apparently valid clones into any self-consistent map were tracked down through case-by-case expert intervention. When the MCD map and the sequence-derived map were compared at the end of the exercise, there was a single discrepancy between them. This discrepancy, which involved an extra EcoRI site in the sequence-derived map, almost certainly reflects a sequencing error since the fusion of the two adjacent sequence-predicted EcoRI fragments was observed in many independently isolated cosmids, all derived from the same yeast strain whose DNA was originally sequenced.

In the course of the resolving inconsistencies, it was discovered that 9/96 of the fragment-size lists contained a single error. The map was based on 957 fragment-size measurements so the overall error rate was <1%. This error rate is still not satisfactory, but two points stand out: all errors were correctly recognized because they led to MCD mapping inconsistencies, and the final map was sufficiently accurate to provide a meaningful quality-control check on the DNA sequence. The latter point is discussed in detail in Gillett et al. (1996). The prediction of a single extra EcoRI site in the sequence-derived map is the error level expected if random sequencing errors occurred in random sequence at a frequency of 0.12%, which is probably fairly typical of the error rates now present in DNA sequence that is produced carefully.

Current results on maps for portions of the HLA Class I region cloned into YACs are less fully analyzed. However, no major new problems have been encountered as we have moved from mapping yeast to human DNA. Much protocol optimization is still needed and major work will be required on the interface between real-world data and the Gillett software package. However, the system is clearly works and one technician could support the mapping demands of a group sequencing approximately Mbp/yr, starting with a YAC contigs. Although the quantitative agreement between fragment sizes in the maps and those predicted by the sequence is not bad (Gillett et al., 1996), we are capable of obtaining better fragment sizing, as documented in Riles et al. (1993). The current experimental system has been in too much flux in recent months to make optimization of fragment sizing a feasible objective. For current applications, fragment-sizing precision is not a critical issue.

B. PLASMID-LEVEL MCD MAPPING

First tests have been carried out of the feasibility of applying MCD mapping at the plasmid level. The basic idea is to substitute restriction enzymes with 4-bp recognition sites for those with 6-bp recognition sites. These enzymes are expected to cleave at approximately 16X the frequency of the 6-bp enzymes and, hence, to produce fingerprints on 2-3-kbp plasmid inserts that have an information content comparable to that of the cosmid fingerprints discussed above. If practiced intensively, plasmid-level mapping could provide a path toward directed sequencing. However, since we are skeptical that any directed-sequencing scheme will provide a cost-effective alternative to deep-shotgun sequencing during the next few years, we consider it more likely, at least in the short term, that plasmid-level MCD mapping will prove useful as an adjunct to deep-shotgun sequencing, aiding with assembly and providing more stringent quality control than would be achievable with cosmid-level maps alone.

An attractive feature of the Gillett software package is its ability to assemble any mixture of sequence contigs and fragment-size lists into MCD maps. The sequence contigs are simply converted to MCD maps before the assembly. Since the MCD-mapping software does not discard any fragment ordering that has already been established, sequence-derived MCD maps enter assembly as fully ordered MCD maps, while individual clones enter assembly as completely unordered MCD maps. The Gillett implementation makes no distinction between the handling of these two types of map.

Current work on plasmid-level MCD mapping as been directed toward establishing a viable experimental system for acquiring the data needed for this style of mapping:

1. Vector Development

We have chosen to proceed initially with a system that entirely parallels the cosmid-level mapping as closely as possible. For this purpose, we need a plasmid vector that is not cleaved by a suitable set of mapping enzymes that have 4-bp recognition sites. Since no such vectors exist, we have a program to create a derivative of the 1.9-kbp plasmid vector, pMOB (Strathmann et al., 1991), that lack sites for the enzymes HinfI, DraI, and RsaI. The two RsaI sites in pMOB were readily eliminated by deleting a short, dispensable fragment in the region between the replication origin and the ampicillin-resistance gene (pMOB's only functional elements). Several HinfI and DraI sites have now also been eliminated by site-direct mutagenesis. A single site for each enzyme remains, one for DraI in the ampicillin-resistance gene and one for HinfI in the replication origin. Despite initial failures, we expect that the former will not be difficult to remove. The latter poses a more difficult problem because of the severe functional constraints on this transcribed-but-not-translated region, which must produce an RNA product that adopts a specific secondary structure.

2. Gel System

We have settled on a composite Metaphor/Seakem agarose gel that produces excellent resolution of fragments ranging in size from approximately 40-4000 bp. These gels are stained with SYBR green, imaged, and analyzed by the same methods employed for the cosmid-level mapping. Some of the parameters embedded in the image-analysis software require different values for the cosmid-level and plasmid-level gel images, but the image-analysis procedure proved robust enough to accommodate this new gel system with no fundamental change.

3. Preliminary Experience

Experience is limited but promising. For historical reasons, an initial cosmid from the human BRCA1 region was chosen for parallel analysis by plasmid-level MCD mapping and shotgun sequencing. The priority of both the mapping and sequencing of this cosmid was lowered after the BRCA1 gene was shown to lie elsewhere. Hence, neither the mapping nor the sequencing has yet been completed and discrepancies between the MCD map and the sequence-derived map have not been resolved. However, preliminary single-digest analysis of the HinfI data suggests that it should be possible to achieve the same--or even better--standards for plasmid-level MCD mapping as for the cosmid-level mapping.

C. RARE Mapping

Long-range continuity in cosmid-level maps is unlikely to be achievable due to biological limitations on the cloning system. While the magnitude of this problem remains unclear, experience from the nematode, where both YAC-level and cosmid-level mapping have been pushed to near exhaustion, is not particularly encouraging (Coulson et al., 1991). Experience with human DNA is less comprehensive but generally consistent with the nematode data: the size of many cosmid contigs appears to be limited to 100-200 kbp by biological factors that are not well understood. Consequently, any comprehensive mapping program must be prepared to orient and align cosmid contigs with lower-resolution maps that have greater continuity.

We have concentrated on methods suitable for aligning cosmid contigs with YAC-based STS-content maps. For this purpose, RARE cleavage (RecA-Assisted-Restriction-Enzyme Cleavage; Ferrin and Camerini-Otero, 1991, 1994; Koob et al., 1992) appears promising. RARE cleavage is an "Achilles Heel" method that allows cleavage with a restriction enzyme to be targeted, via the homology-recognizing specificity of RecA-oligodeoxynucleotide complexes, to a particular member of a large set of sites. We have gained considerable experience cleaving YACs with EcoRI by RARE-targeting to a specific EcoRI site. For example, this method allows precise length calibration of maps based on overlapping YACs since RARE cleavage of one YAC at the EcoRI site that corresponds to the appropriate vector-insert junction of its overlap partner defines the phasing between the two YACs (Gnirke et al., 1994). We have also succeeded in generating specific EcoRI partial-digest fragments by double-RARE cleavage of mammalian DNA (Gnirke et al., 1993); however, we find this procedure to be insufficiently robust for routine application.

Note: A manuscript published subsequent to the completion of this project (Wong et al., 1997) provides detailed documentation of the success of large-scale applications of cosmid-level MCD mapping. Most of the R&D behind this first large-scale test of the system was supported under DE-FG06-92ER61487.

REFERENCES

Agard, D.A., Steinberg, R.A., and Stroud, R.M. (1981). Quantitative analysis of electrophoretograms: a mathematical approach to super-resolution. Anal. Biochem. 111: 257-268.

Birren, B., Hood, L., and Lai, E. (1989). Pulsed field gel electrophoresis: studies of DNA migration made with the programmable, autonomously-controlled electrode electrophoresis system. Electrophoresis 10: 302-309.

Carle, G.,F., Frank, M., and Olson, M.V. (1986). Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. Science 232: 65-68.

Carrano, A.V., Lamerdin, J., Ashworth, L.K., Watkins, B., Branscomb, E., Slezak, T., Raff, M., de Jong, P.J., Keith, D., McBride et al. (1989). A high resolution fluorescence-based, semiautomated method for DNA fingerprinting. Genomics 4: 129-136.

Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J., and Waterston, R. (1991). YACs and the C. elegans genome. Bioessays 13: 413-417.

Coulson, A., Waterston, R., Kiff, J., Suston, J., and Kohara, Y. (1988). Genome linking with with yeast artificial chromosomes. Nature 335: 184-186.

Craig, A.G., Nizetic, D., Hoheisel, J.D., Zehetner, and G., Lehrach, H. (1990). Ordering of cosmid clones covering the herpes simplex virus tyep I (HSV-1) genome: a test case for fingerprinting by hybridisation. Nucleic Acids Res. 18: 2653-2660.

Drury, H.A., Clark, K.W., Hermes, R.E., Feser, J.M., Thomas, L.J., Jr., and Donis-Keller, H. (1992). A graphical user interface for quantitative imaging and analysis of electrophoretic gels and autoradiograms. Biotechniques 12: 892-898, 900-901.

Drury, H.A., Green, P., McCauley, B.K., Olson, M.V., Politte, D.G., and Thomas, L.J.,Jr. (1990). Spatial normalization of one-dimensional electrophoretic gel images. Genomics 8: 119-126.

Evans, G.A., Lewis, K., and Rothenberg, B.E. (1989). High efficiency vectors for cosmid microcloning and genomic analysis. Gene 79: 9-20.

Ferrin, L.J., and Camerini-Otero, R.D. (1991). Selective cleavage of human DNA: RecA-assisted restriction endonuclease (RARE) cleavage. Science 254: 1494-1497.

Ferrin, L.J., and Camerini-Otero, R.D. (1994). Long-range mapping of gaps and telomeres with RecA-assisted restriction endonuclease (RARE) cleavage. Nat. Genet. 6: 379-383.

Galat, A., and Goldberg, I.H. (1987). Analysis of microdensitometric data in terms of probability of cleavage of DNA. Comput. Appl. Biosci. 3: 333-338.

Geraghty, D. (1993). Structure of the HLA class I region and expression of its resident genes. Curr. Opin. Immunol. 5: 3-7.

Geraghty, D.E., Pei, J., Lipsky, B., Hansen, J.A., Taillon-Miller, P., Bronson, S.K., and Chaplin, D.D. (1992). Cloning and physical mapping of the HLA class I region spanning the HLA-E to HLA-F interval by using yeast artificial chromosomes. Proc. Natl. Acad. Sci. USA 89: 2669-2673.

Gillett, W., Hanks, L., Wong, G.-K-S., 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.

Gnirke, A., Huxley, C., Peterson, K., and Olson, M.V. (1993). Microinjection of intact 200- to 500-kb fragments of YAC DNA into mammalian cells. Genomics 15: 659-667.

Gnirke, A., Iadonato, S.P., Kwok, P.-Y., and Olson, M.V. (1994). Physical calibration of yeast-artificial-chromosome-contig maps by RecA-assisted restriction endonuclease (RARE) cleavage. Genomics 24, 199-210.

Graham, M.Y., Otani, T., Boime, I., Olson, M.V., Carle, G.F., and Chaplin, D.D. (1987). Cosmid mapping of the human chorionic gonadotropin beta subunit genes by field-inversion gel electrophoresis. Nucleic Acids Res. 15: 4437-4448.

Gray, A.J., Beecher, D.E., and Olson, M.V. (1984). Computer-based image analysis of one-dimensional electrophoretic gels used for the separation of DNA restriction fragments. Nucleic Acids Res. 12: 473-491.

Helms, C., Dutchik, J.E., and Olson, M.V. (1987). A lambda DNA protocol based on purification of phage on DEAE-cellulose. Methods Enzymol. 153: 69-82.

Kohara, Y., Akiyama, K., and Isono, K. (1987). The physical map of the whole E. coli chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. Cell 50: 495-508.

Koob, M., Burkiewics, A., Kur, J., and Szybalski, W. (1992). RecA-AC: single site cleavage of plasmids and chromosomes at any predetermined restriction site. Nucleic Acids Res. 20: 5831-5836.
Lai, E., Wang, K., Avdalovic, H., and Hood, L. (1991). Rapid restriction map constructions using a modified pWE15 cosmid vector and a robotic workstation. Biotechniques 11: 212-214.

Lamerdin, J.E., and Carrano, A.V. (1993). Automatic fluorescence-based restriction fragment analysis. Biotechniques 15: 294-303.

Lee, L.G., Chen, C.H., and Chiu, L.A. (1986). Thiazole orange: a new dye for reticulocyte analysis. Cytometry 7: 508-517.

Mizukami, T., Chang, W.I., Garkavtsev, I., Kaplan, N., Lombardi, D., Matsumoto, T., Niwa, O., Kounosu, A., Yanagida, M., Marr, T.G. et al. (1993). A 13 kb resolution cosmid map of the 14 Mb fission yeast genome by nonrandom sequence-tagged site mapping. Cell 73: 121-132.

Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., Frank, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. Proc. Natl. Acad. Sci. USA 8: 7826-7830.

Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., and Frank, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. Proc. Natl. Acad. Sci. USA 83: 7826-7830.

Pieretti, M., Tonlorenzi, R., and Ballabio, A. (1991). Rapid assembly of lambda phage contigs within YAC clones. Nucleic Acids Res. 19: 2795-2796.

Rackwitz, H.R., Zehetner, G., Frischauf, A.M., and Lehrach, H. (1984). Rapid restriction mapping of DNA cloned in lambda phage vectors. Gene 30: 195-200.

Riles, L., Dutchik, J.E., Baktha, A., McCauley, B.K., Thayer, E.C., Leckie, M.P., Braden, V.V., Depke, J.E., Olson, M.V. (1993). Physical maps of the six smallest chromosomes of Saccharomyces cerevisiae at a resolution of 2.6 kilobase pairs. Genetics 134: 81-150.

Riles, L., Dutchik, J.E., Baktha, A., McCauley, B.K., Thayer, E.C., Leckie, M.P., Braden, V.V., Depke, J.E., and Olson, M.V. (1993). Physical maps of the six smallest chromosomes of Saccharomyces cerevisiae at a resolution of 2.6-kilobase pairs. Genetics 134: 81-150.

Smith, H.O., and Birnstiel, M.L. (1976). A simple method for DNA restriction site mapping. Nucleic Acids Res. 3: 2387-2398.

Smith, J.M., and Thomas, D.J. (1990). Quantitative analysis of one-dimensional gel electrophoresis profiles. Comput. Appl. Biosci. 6: 93-99.

Stallings, R. L., Doggett, N. A., Okumura, K., and Ward, D. C. (1992). Chromosome 16-specific repetitive DNA sequences that map to chromosomal regions known to undergo breakage/

Stallings, R.L., Torney, D.C., Hildebrand, C.E., Longmire, J.L., Deaven, L.L., Jett, J.H., Doggett, N.A., and Moyzis, R.K. (1990). Physical mapping of human chromosomes by repetitive sequence fingerprinting. Proc. Natl. Acad. Sci. USA 87: 6218-6222.

Strathmann, M., Hamilton, B.A., Mayeda, C.A., Simon, M.I., Meyerowitz, E.M., and Palazzolo, M.J. (1991). Transposon-facilitated DNA sequencing. Proc. Natl. Acad. Sci. USA 88: 1247-1250.

Takahashi-Fuji, A., Maeda, N., Ishino, Y., Kotani, H., and Kato, I. (1994). Non-iosotopic restriction mapping of cosmid DNA. Biotechniques 16: 910-915.

Whittaker, P.A., Wood, L., Mathrubutham, M., and Anand, R. (1993). Generation of ordered phage sublibraries of YAC clones: construction of a 400-kb phage contig in human dystrophin gene. Genomics 15: 453-456.

Wong, G.K.-S., Yu, J., Thayer, E.C., and Olson, M.V. (1997). Multiple-complete-digest (MCD) restriction-fragment mapping: Generating sequence-ready Maps for large-scale DNA sequencing. Proc. Natl. Acad. Sci. USA 94: 5225 - 5230.