Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 7821–7825, October 1986
Genetics

Toward a physical map of the genome of the nematode *Caenorhabditis elegans*

(ordered clone bank/genomic data base/clone matching)

ALAN COULSON, JOHN SULSTON, SYDNEY BRENNER, AND JONATHAN KARN

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, England

Contributed by Sydney Brenner, June 23, 1986

ABSTRACT **A technique for digital characterization and comparison of DNA fragments, using restriction enzymes, is described. The technique is being applied to fragments from the nematode *Caenorhabditis elegans* (*i*) to facilitate cross-indexing of clones emanating from different laboratories and (*ii*) to construct a physical map of the genome. Eight hundred sixty clusters of clones, from 35 to 350 kilobases long and totaling about 60% of the genome, have been characterized.**

We are engaged in the construction of a physical map of the genome of the nematode *Caenorhabditis elegans*. The map will ultimately consist of a fully overlapping collection of cloned DNA fragments, insofar as this can be achieved in a reasonable period of time. The fragments are held permanently as frozen cosmid or λ clones, and restriction digest data characteristic of each is placed in a computer data base so that incoming clones can be compared with old ones. The project is of necessity lengthy, and is as yet far from completion, but we hope that our experiences to date will be of some interest. Olson (1) describes a parallel project, using different methodology, on the genome of *Saccharomyces cerevisiae*.

At approximately 8×10^7 base pairs (bp) (2), the genome of *C. elegans* is the smallest known for any metazoan. The genetic map carries some 500 known loci, and there is a large and continually increasing set of cloned genes, restriction fragment length polymorphisms, and genetic breakpoints by which the genetic and physical maps can be correlated. Many clones have been localized to chromosomal regions by *in situ* hybridization (3).

The major benefit of the physical map will be immediate access to any segment of the genome that can be defined genetically. Additionally, it will be a starting point for studying the large-scale organization of the genome. However, there are two intermediate goals on the way toward the creation of the full map. The first and most important is to provide communication between the various laboratories engaged in cloning segments of the *C. elegans* genome. The second is the provision of flanking sequences when a segment is found to match fragments already in the data base.

Given these priorities, our first step was the choice of a suitable restriction "fingerprinting" procedure for matching clones to one another so that overlaps could be recognized. Mapping is now proceeding in two stages. In the first, clones are picked at random and compared with one another to yield a mixture of contigs† (i.e., groups of clones with contiguous nucleotide sequences) and unattached clones. In the second, clones will be preselected by means of hybridization probes taken from the ends of contigs and from unattached clones, to fill the gaps.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Bacteria. pJB8 cosmid recombinants were grown in *Escherichia coli* 1046 (4). LoristB cosmids were grown in ED8767 (5). λ 2001 recombinants were grown in Q358 (6).

Vectors. Cosmid pJB8 was as described by Ish-Horowicz and Burke (7). The cosmid loristB, a modification of loric (8), was a gift of P. Little. λ 2001 was as described by Karn *et al.* (6).

Enzymes and Chemicals. Avian myeloblastosis virus reverse transcriptase and 60 units/μl *Sau*3A1 were purchased from Anglian Biotechnology Ltd. (Colchester, England). Other enzymes were from New England Biolabs. [α-^{32}P]dATP was from Amersham. γ-Methacryloxypropyltrimethoxysilane was from Wacker Chemie (Munich, F.R.G.).

Isolation of *C. elegans* DNA. *C. elegans* (Bristol Laboratories, Syracuse, NY) was grown in liquid culture (2). A 1-g aliquot of frozen nematodes was ground to a powder under liquid nitrogen, mixed gently into 30 ml of lysis buffer (100 mM EDTA, pH 8/5 mM Tris·HCl/0.5% NaDodSO$_4$/proteinase K at 50 μg/ml), incubated at 50°C for 2 hr, and gently extracted with cold phenol. Nucleic acids were precipitated with ethanol and dispersed in TE (10 mM Tris·HCl/0.5 mM EDTA, pH 8).

Recombinant Constructions. Randomly selected clones from a variety of different libraries have been analyzed in this study. Partial *Sau*3A1, *Mbo* I, and *Eco*RI fragments were inserted into pJB8, and partial *Mbo* I fragments were inserted into loristB. We have also analyzed clones from an *Mbo* I digest inserted into pHC79 (G. Benian and R. Waterston, personal communication) and gridded at MIT (G. Ruvkun and R. Horvitz, personal communication); from a λ2001 library constructed here; and many individual clones that were received from other *C. elegans* laboratories.

Cosmid Banks. Four 20-μg aliquots of freshly extracted nematode DNA were partially digested with various concentrations of restriction enzyme in the presence of ribonuclease, mixed, and loaded onto a 0.4% LGT agarose (SeaKem Laboratories, Rockland, ME) gel. The required size fraction [30–50 kilobases (kb)] was cut out and repurified on a second gel. About 1 μg of sized DNA was recovered. Of this DNA, 0.2 μg was ligated to 0.1 μg of each of the *Hin*dIII/*Eco*RI and *Sal* I/*Eco*RI arms of pJB8 as described by Ish-Horowicz and Burke (7). Packaging of the ligated DNA, with mixes derived from *E. coli* strains NS428 and NS433 (9), yielded up to 10^6 recombinants per μg of insert DNA. The partial *Sau*3A1 library was prepared by cloning the fragments into an excess of *Bam*HI dephosphorylated vector, giving a yield of about 3×10^4 recombinants per μg of insert DNA.

Abbreviations: kb, kilobase(s); contig, a group of cloned nucleotide sequences that are contiguous; bp, base pair(s).

†The term "contig" was introduced by Rodger Staden (20) in connection with DNA sequence analysis.

Following adsorption onto *E. coli* 1046 and plating at low density on ampicillin (50 μg/ml) plates, the cosmids were transferred to microtiter plate wells containing 2× TY medium with ampicillin at 75 μg/ml [cf. Gergen *et al.* (10)] and incubated at 37°C for 24 hr. The suspensions were made about 20% (vol/vol) in glycerol and then carefully stirred with a 96-prong "hedgehog" made from 3-mm diameter brass rods set in a plastic block. This device was also used for replication of the bank onto agar plates or into secondary microtitre plates. Between operations, the hedgehog was rinsed under running water, soaked for about 30 sec in ethanol, and flamed. The microtiter plates were stored in polythene bags at −70°C.

λ Banks. λ banks were prepared from DNA sized on gels in a similar way and ligated into the vector λ2001 (6, 11).

Cosmid Minipreps. Cosmid DNA was extracted essentially by the alkaline $NaDodSO_4$ method of Birnboim and Doly (12) as modified by Ish-Horowicz and Burke (7). Generally, 48 or 96 preparations were carried out simultaneously, starting from 2-ml portions of cultures grown to saturation in 2× TY containing ampicillin (75 μg/ml) [or kanamycin sulfate (30 μg/ml) for loristB cosmids]. After two ethanol precipitations, the product (1–3 μg) was dispersed in 30 μl of TE.

Fingerprinting of Cosmid DNA. To analyze 48 clones, 100 μl of a reaction mixture was made containing 20 μCi of [^{32}P]dATP (1 Ci = 37 GBq) (4×10^5 Ci/mol), 2× medium salt restriction endonuclease buffer (11), 20 μg of boiled RNase, 25 μM dideoxyGTP, 40 units of *Hin*dIII endonuclease, and 40 units of avian myeloblastosis virus reverse transcriptase. Two microliters of this mix was aliquoted into each of 48 wells of a 10-μl well microtitre plate (NUNC1-6311B 60 × 10 μl) precooled on ice, using a Hamilton PB600-1 repetitive dispenser fitted with a siliconized disposable tip. Miniprep DNA (see above) (1–1.5 μl) was added to each well. The reactions were sealed with a glass plate covered in Parafilm and incubated for 45 min at 37°C. The reverse transcriptase was inactivated by a 30-min incubation at 68°C. The plate was cooled on ice, and 2 units of *Sau*3A1 in 2 μl of 1× restriction buffer (11) were added to each well. The plate was resealed and incubated for 2 hr at 37°C. The reactions were terminated by the addition of 4 μl of 98% (vol/vol) formamide/0.3% bromophenol blue/0.3% xylene cyanol/10 mM EDTA. Just prior to gel loading the samples were denatured at 80°C for 10 min.

The fragments for the marker lanes on the gels were "end filled" *Sau*3A1 digests of λDNA, similarly denatured. The 20 × 40 cm × 0.35 mm denaturing polyacrylamide gels were 4% acrylamide/bis-acrylamide (19:1, wt/wt)/8 M urea in TBE (13, 14). The gel was bonded to one plate with methacryloxypropyltrimethoxysilane by the method of Garoff and Ansorge (15) to prevent distortion of the wells prior to sample loading. A 3-mm thick aluminum plate was clamped to the front of the gel to minimize "smiling" by improving heat distribution. The fingerprint reaction mixtures (3 μl) were loaded, interspersed every 6 lanes with 1 μl of marker digest. Electrophoresis was for 1.75 hr at 30 W. After fixing in 10% (vol/vol) acetic acid for 15 min, washing in tap water for 30 min, and drying, the gels were autoradiographed, without an intensifying screen, for 2–3 days.

Computer Programs. Programs were written in Fortran 77 and run on a DEC Vax 8600 computer using a VMS operating system. They are available on request.

Digitization. A Grafbar digitizer is used; the outputs are clone name, gel identifier, and band coordinates.

Matching. For each clone pair the number of bands that agree within a preset tolerance (typically 0.7–1.0 mm) is recorded, and the probability of this event occurring by chance is calculated from the number of bands in each clone and the tolerance. All matches for each incoming clone are ranked in terms of this probability, and the first few are printed out. The output includes additional information about the positions of the matching clones in contigs.

Assembly. Currently, clones that are seen to lie internally within contigs can be semi-automatically entered in "*" format (see below) by the computer; others are handled interactively with an editing program. Contigs are displayed (see Fig. 3), and the operator uses a cursor to manipulate clones and contigs (adding, positioning, and deleting clones; joining contigs).

Retrospective searches. Other programs are used to hunt for matches among clones that lie at the ends of contigs or are as yet unattached. Possible matches are ranked by examination of overlapping clones for logical fits, and the most plausible are printed out for further evaluation. This approach becomes increasingly useful as the project develops.

Statistics. We have experimented with a variety of analytical techniques for assessing progress. The two most useful are the log (run regularly to generate data for the progress curve), and the histogram of band occurrence (see *Results and Discussion*).

Models. Fictitious data bases are generated either by purely random assembly or by starting from the known properties of the real map to date.

RESULTS AND DISCUSSION

Fingerprinting Method. The method used for fingerprinting the clones is shown in Fig. 1. Cloned DNA is digested by a restriction enzyme with a 6-bp specificity that leaves staggered ends, and simultaneously the ends are labeled by end-filling with reverse transcriptase and a suitable mixture of triphosphates. The inclusion of a dideoxy triphosphate is desirable to ensure a reproducible extent of filling.

Next, the enzymes are destroyed by heat, and the DNA fragments are cleaved again, this time by a restriction enzyme with a 4-bp specificity. The resulting small fragments are of a size suitable for separation on a thin polyacrylamide gel. We find that denaturing gels give resolution superior to that of nondenaturing gels. The combination of *Hin*dIII and *Sau*3A1 has proved satisfactory for nematode cosmids, yielding about 23 bands on average and allowing unambiguous assignments for overlaps of one-third to one-half of the bands.

If the 4-bp specific enzyme has the same cleavage specificity as the enzyme used to make the bank, there will be no anomalously sized bands resulting from fusion of a *C. elegans* fragment with a vector fragment. We have taken advantage of this feature for most of our work. In practice, however,

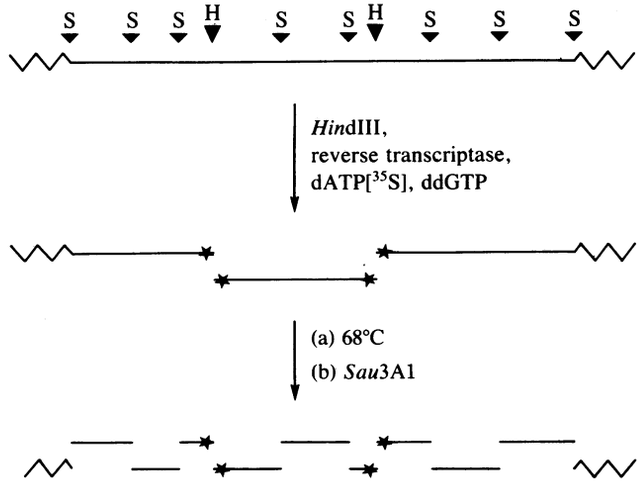

FIG. 1. Fingerprinting procedure. ddGTP, dideoxyGTP; S, *Sau*3A1 sites; H, *Hin*dIII sites.

such end effects are never a serious problem for cosmid clones, because the final fragments are small and numerous.

The gels (Fig. 2) are calibrated by means of marker lanes (containing λ DNA totally digested with *Sau*3A1 and end-labeled). In our previous, manual, technique for data entry, a digitizing tablet was provided with a grid of lines drawn to correspond with a canonical marker lane. The position of the grid was communicated to the computer by entry of fiducial marks, and then the film was locally aligned with the grid as digitization of the sample bands proceeded. We have begun to use a digitizing scanner that allows semi-automatic data entry, with improvement in accuracy and speed (J. F. Mallett and J.S., unpublished data).

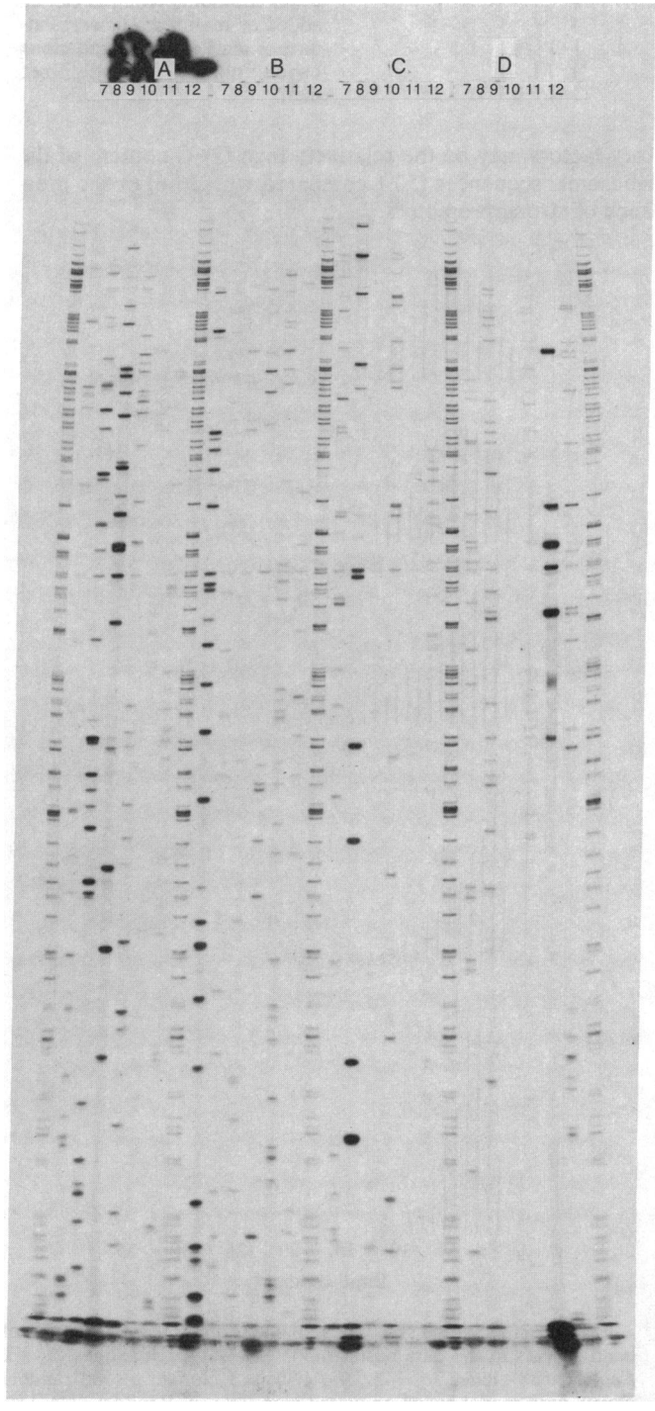

FIG. 2. Autoradiograph of a typical mapping gel. The five lanes with closely spaced bands are markers: fragments of size 58–2225 bases can be seen. Sample D11 is ribosomal.

Contig Assembly. Each clone is now compared by the computer with the entire data base, and a rank order of the most likely matches is printed out, together with additional information.

The computer does not actually assemble the contigs. This is done by an interactive program, so that we can ourselves judge the reliability of each match by direct comparison of the films. Visual alignment provides much more precision than is available in the digitized data. The program used for assembly has a variety of routines that can be called as required for making subsidiary comparisons and for annotating the clones.

Fig. 3 shows the computer screen during contig assembly. The lengths of the lines are proportional not to kilobases but to number of bands (or, roughly, to number of *Hin*dIII sites). Length calculations are based on the mean size of insert—found to be about 34 kb for the cosmids. "*" signifies that additional similar clones are in the data base but are not displayed; redundant clones are buried in this way to avoid excessive clutter on the screen.

So far, there have been no obvious ambiguities resulting from repeated DNA sequences. Evidently the 10-bp specificity of each band in the fingerprint ensures that most dispersed repeats are either too short or too inaccurate to be detected in the context of a cosmid or λ clone. No homology is detected between the members of gene families [e.g., myosin (16), vitellogenin (17), collagen (18), and major sperm protein (19)]. Large accurate tandem repeats will yield relatively abundant clones carrying fragments from within the repeat structure, together with rare clones carrying end fragments. Our one known example is the ribosomal cluster (21, 22), though the statistics are complicated by the superior viability of ribosomal clones (see below). Small accurate tandem repeats containing *Hin*dIII sites are detected by the appearance of heavy bands in the fingerprint, and indeed two clones that yield such fingerprints have been shown to give ladder patterns on agarose gels after partial *Hin*dIII digestion.

Assessment of Progress. The progress of the project is illustrated in Fig. 4. The percentage scale in Fig. 4*a* is subject to the uncertainty in the estimate of genome size made by Sulston and Brenner (2), but the slope of the line showing the actual length of DNA in contigs can be compared directly and pragmatically with what we can achieve by prior selection of clones.

To assess the situation objectively, we have made calculations of the progress expected if our clones were randomly distributed in the genome (thin lines in Fig. 4 *a* and *b*). For this purpose, we create model data bases using the computer's random number generator. The progress curve lies below that predicted by the models, indicating that the banks we have used are not perfectly random. This can be seen by the initial excess of matches over that expected from a random system (Fig. 4*a*).

To compare different banks with one another we use the assay technique shown in Fig. 5. In the histograms each bar represents the number of bands that occur a given number of times in the data base. Bands that occur only once because they are in unattached clones are plotted at U, separately from bands that occur only once because they are in clones projecting from the ends of contigs. It is at once apparent that the real data bases (Fig. 5 *a–c*) contain more bands at high repetition frequencies than does the model (Fig. 5*e*)—in other words, the clones are distributed less evenly through the genome than would be expected by chance. To allow realistic assessment, the result of adding truly random clones to our existing data base has been modeled by another program (Fig. 5*d*).

We tentatively conclude from the comparison of the *Eco*RI bank with the *Sau*3A1 bank that the lack of randomness does not arise from the uneven distribution of restriction sites in

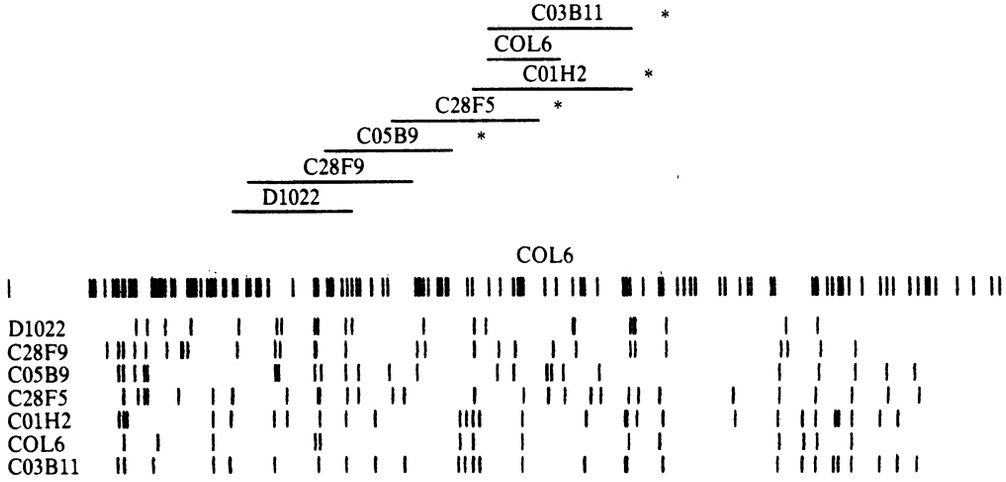

FIG. 3. A contig displayed on the computer screen. (*Upper*) Each clone is represented by a line of length proportional to the number of bands. Asterisk indicates the presence of hidden clones. Repeat of the name COL6 beneath the contig indicates location of this known gene; additional remarks can be added as required. (*Lower*) Pattern of marker bands and clone bands, plotted from digitized data.

the genome. Rather, it seems to result from the relative viability of cosmids containing different segments of nematode DNA. Curiously, segments of ribosomal DNA clone remarkably well—up to 5% of clones in the primary cosmid banks are ribosomal, compared with the 0.05% expected (2, 22). The same phenomenon is seen in amplified λ banks but not in primary λ banks. Why the ribosomal DNAs should confer this selective advantage is not known. One possible component may be the absence of *Eco*K sites from the ribosomal sequences, since *Eco*K activity has been demonstrated in standard packaging extracts (23). Other contributory factors may be the relatively high G+C content of the ribosomal sequences (50% compared with 36%) or the presence of strong promoters.

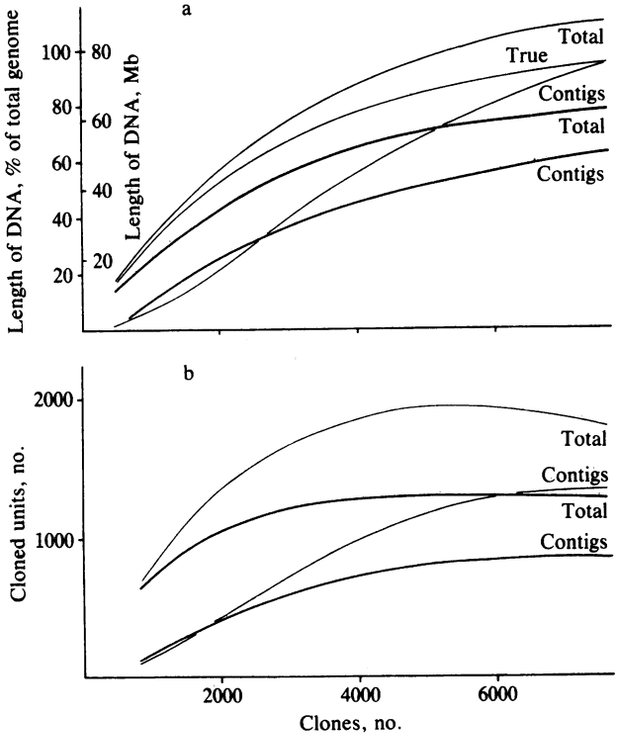

FIG. 4. Progress curves. Thick lines represent experimental data. Thin lines represent model data, generated on the basis of a genome size of 8×10^7 bp and a cloned insert size of 3×10^4 bp (although the cosmid inserts are somewhat larger than this on average, we use this value to allow for the presence of some λ clones in the data base) with a minimum detectable overlap of 50%. (*a*) Estimated length of DNA cloned into contigs and into total of contigs plus unattached clones. The "true" curve for the model data differs from the "total" curve in that it takes into account undetected overlaps. Note the absence of an initial dip in the experimental contig length curve: this is a direct indication of nonrandom cloning. Mb, megabases. (*b*) Number of cloned units. "Total" refers to contigs plus unattached clones.

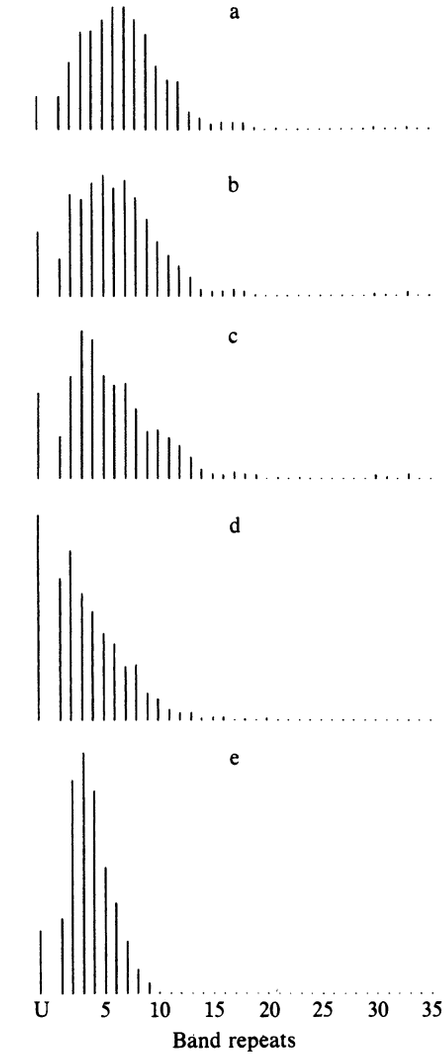

FIG. 5. Histograms showing actual and predicted distribution of band repeats in the data base. *a*, 593 *Sau*3A1/pJB8 clones. *b*, 870 *Eco*RI/pJB8 clones. *c*, 342 *Mbo* I/loristB clones. *d*, 500 random clones onto model based on actual map data. *e*, Random data base (minimum overlap 0.5). Ribosomal clones are excluded from the analysis. The progress curves in Fig. 4 are derived from the same data.

Apart from the slight advantage conferred by the loristB vector, we have no direct way of overcoming the nonrandomness of the banks. Equally, however, there is no evidence as yet for segments of DNA that cannot be cloned at all. The redundancy of the clones merely means that more of them have to be examined for a given advance in the map.

The mean contig size is at present 56 kb, in close agreement with the size predicted by the random model (58 kb). There are 16 contigs of length greater than 150 kb, the largest being 360 kb, but 6 of these depend upon gaps being filled by preselection of clones. The largest contig attained by random assembly is 250 kb in size, again in good agreement with the model (230 kb—though, of course, this degree of precision is fortuitous).

CONCLUSION

We have achieved our first goal of establishing "genomic communication" among *C. elegans* laboratories. Of the clones that we have received, about two-thirds have been placed in contigs. In a number of places, physical linkage between pairs of such clones has thereby been established.

The nematode genome probably lies near the upper limit of size for matching by single-lane fingerprinting. This size class would include individual mammalian chromosomes. However, with some automation, the existing method could be extended to the entire genome of a higher eukaryote, provided that two or more lanes, each using a different pair of enzymes, were electrophoresed for each clone. Compared with the method of Olson (1), our method has the advantages of more sensitive pairwise matching and greater tolerance for clones from diverse sources, but it has the disadvantage of not directly generating a restriction map.

Progress towards the complete nematode map now depends upon the efficiency with which the missing pieces can be found and will increasingly become a communal effort as the "genomic walks" carried out in the various *C. elegans* laboratories are brought together into a common reference data base. Full connection will take a considerable time to achieve and indeed may not be practicable at all. Even near completion, however, the map will be of great value both for studying the large scale organization of the genome and for isolating and characterizing segments that cannot be readily identified in other ways.

We would first like to thank *C. elegans* researchers everywhere for the practical support that they have given to us in undertaking this project; without their cooperation the work would have little purpose. We are indebted to Maynard Olson for stimulating discussions; to Rodger Staden for writing the original programs used for digitization and matching; to Iva Greenwald for her help in setting up the first cosmid bank; to Peter Little for the provision of lorist vectors; to Guy Benian and Gary Ruvkun for exchanging cosmid clone banks with us. For additional advice and discussion we are grateful to numerous colleagues—in particular Donna Albertson, George Brownlee, Ronald Ellis, Toby Gibson, Jonathan Hodgkin, Bob Horvitz, Sam Ward, and Bob Waterston.

1. Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. & Frank, T. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7826–7830.
2. Sulston, J. E. & Brenner, S. (1974) *Genetics* **77,** 95–104.
3. Albertson, D. G. (1985) *EMBO J.* **4,** 2493–2498.
4. Cami, B. & Kourilsky, P. (1978) *Nucleic Acids Res.* **5,** 2381–2390.
5. Murray, N. E., Brammar, W. J. & Murray, K. (1977) *Mol. Gen. Genet.* **150,** 53–61.
6. Karn, J., Matthes, H. W. D., Gait, M. J. & Brenner, S. (1984) *Gene* **32,** 217–224.
7. Ish-Horowicz, D. & Burke, J. F. (1981) *Nucleic Acids Res.* **9,** 2989–2998.
8. Little, P. F. R. & Cross, S. H. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 3159–3163.
9. Sternberg, N., Tiemeier, D. & Enquist, L. (1977) *Gene* **1,** 255–280.
10. Gergen, J. P., Stern, R. H. & Wensink, P. C. (1979) *Nucleic Acids Res.* **7,** 2115–2136.
11. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
12. Birnboim, H. C. & Doly, J. (1979) *Nucleic Acids Res.* **7,** 1513–1523.
13. Peacock, A. C. & Dingman, C. W. (1968) *Biochemistry* **7,** 668–674.
14. Sanger, F. & Coulson, A. R. (1978) *FEBS Lett.* **87,** 107–110.
15. Garoff, H. & Ansorge, W. (1981) *Anal. Biochem.* **115,** 450–457.
16. Karn, J., Dibb, N. J. & Miller, D. M. (1985) in *Cell and Muscle Motility*, ed. Shay, J. W. (Plenum, New York), Vol. 6, pp. 185–237.
17. Blumenthal, T., Squire, M., Kirtland, S., Cane, J., Donegan, M., Spieth, J. & Sharrock, W. (1984) *J. Mol. Biol.* **174,** 1–18.
18. Cox, G. N., Carr, S., Kramer, J. M. & Hirsh, D. (1985) *Genetics* **109,** 513–528.
19. Klass, M. R., Kinsley, S. & Lopez, L. C. (1984) *Mol. Cell Biol.* **4,** 529–537.
20. Staden, R. (1980) *Nucleic Acids Res.* **8,** 3673–3694.
21. Files, J. G. & Hirsh, D. (1981) *J. Mol. Biol.* **149,** 223–240.
22. Ellis, R. E., Sulston, J. E. & Coulson, A. R. (1986) *Nucleic Acids Res.* **14,** 2345–2364.
23. Rosenberg, S. M. (1985) *Gene* **39,** 313–315.