LBL-29093

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

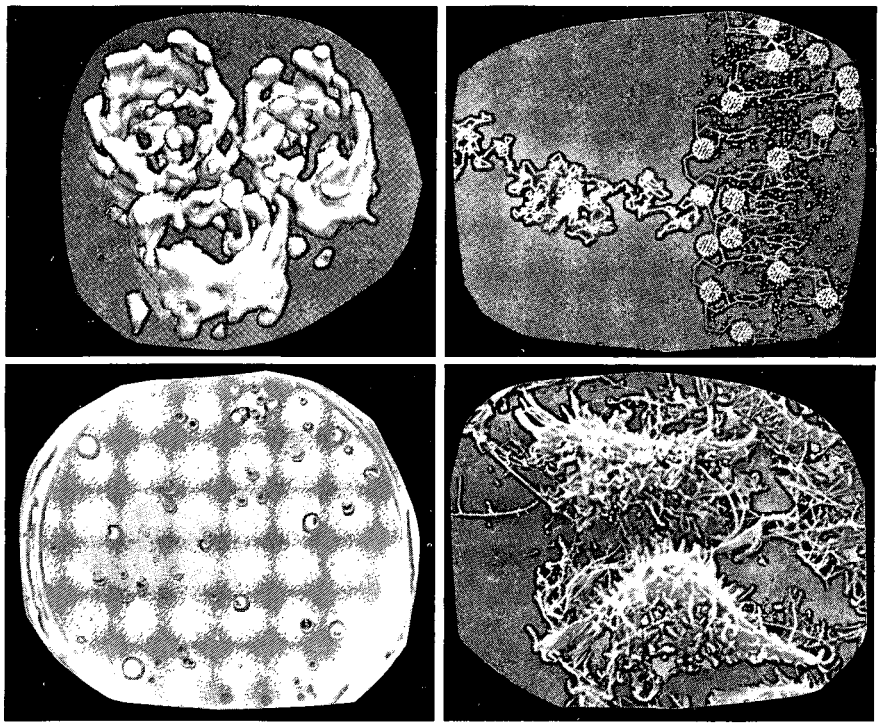
CELL & MOLECULAR BIOLOGY DIVISION

To be published as a chapter in **Structure and Function of Nucleic Acids and Proteins**, F. Wu and C.-W. Wu, Eds., Raven Press Publisher, NY, NY, 1990

Challenges in the Human Genome Project

C.R. Cantor and C.L. Smith

April 1990



LOAN COPY
Circulates
for 2 weeks
Bldg. 50 Library.
LBL-29093
Copy 2

Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal 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 its 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, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

CHALLENGES IN THE HUMAN GENOME PROJECT

Charles R. Cantor* and Cassandra L. Smith+

Departments of Genetics and Development,* Microbiology+ and Psychiatry,+ College of Physicians and Surgeons, Columbia University, New York, NY 10032#

ABSTRACT

The human genome project is designed with the assumption that it will be possible to achieve very substantial improvements in the rate and efficiency of mapping, sequencing, and data interpretation. Here a number of the technological obstacles which must be overcome will be outlined and possible strategies for circumventing these obstacles will be discussed. Some obstacles derive just from the enormous scale of the human genome and the relatively high degree of DNA polymorphism in the human population. Others are the result of intrinsic limitations in current mapping and sequencing strategies, and limitations in our current understanding of recombination, gene structure, and protein function. The advances we will have to make to complete the genome project will, in overcoming the latter obstacles, provide new biological insights that should have implications far beyond the project itself.

INTRODUCTION

The goals of the human genome project are to complete several different kinds of physical maps, to increase the resolution of the existing genetic map, and to develop the tools needed to locate and identify the estimated 50,000 to 100,000 human genes. These goals could not be easily accomplished simply by utilizing currently available technology. Thus the genome project is projected to occur in a climate of evolving technology. For the project to succeed, one to two orders of magnitude improvements will be needed in such areas as sequencing speed

#Current address: Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720

and cost, sample archive management, and data acquisition and distribution. For example, until very recently, most schemes for large scale DNA sequencing relied on pre-existing dense clone banks covering the region of interest. It is not at all clear that such samples can be constructed for entire complex genomes. As another example, while most protein coding regions can be identified by inspection of DNA sequences, it is not clear that small exons can be found reliably.

Here a number of the technical and methodological obstacles currently facing the human genome project will be discussed, and some possible solutions will be described, where these can be envisaged as reasonable extrapolations from existing methods.

FINISHING MAPS

Two basic approaches dominate current physical mapping efforts. These are usually referred to as top down and bottom up mapping. In the former approach, ideally one starts with a single chromosome, and the goal is a restriction map. The chromosomal DNA is subdivided into a discrete number of specific fragments by cleavage with a restriction enzyme with very rare cutting sites, and fractionation of the resulting mixture by pulsed field gel electrophoresis (PFG). Most of these fragments are placed in order by ordinary Southern analysis using cloned DNA probes that have been assigned approximate or relative positions along the chromosome either by a pre-existing genetic map, or by in situ hybridization or somatic cell genetics. For fragments where no corresponding cloned DNA exists, one can infer the order by the analysis of partial digests, probed from neighboring fragments, or by using specialized probes such as linking clones which can often determine the order of two fragments based solely on their size.

In bottom up mapping, the goal is an ordered library of cloned fragments such as bacteriophages, cosmids, or yeast artificial chromosomes,YACs (1-3). In the future, given the potential power of the polymerase chain reaction (PCR) it may not necessary to possess the actual clones as long as one has the information needed to amplify their DNA through the use of the appropriate primers (4). Clones are randomly selected and fingerprinted to determine bits of sequence information. In most approaches this is actually partial restriction map information. As

increasing numbers of clones are studied, pairs and then multiples with overlapping sequence information are detected. These form contiguous blocks of cloned DNA (contigs). Ideally the blocks will eventually coalesce to form a complete ordered library which can then be simplified by choosing those clones that have the minimum overlap so that the genome is covered with a minimum clone set.

In practice, with both general approaches, it is very easy to start maps and very hard to finish them (Fig. 1). At first, with either approach, every DNA segment or clone selected is interesting and adds to the available map information. However as the project progresses, restriction mapping slows down because the smaller DNA fragments are unlikely, a priori, to be represented in cloned DNA and thus they are more difficult to place on the map. Library ordering slows down because most new clones selected at random will cover ground already mapped. In practice one can complete restriction mapping efforts by usually specialized strategies that focus on the smaller fragments. In organisms without extensive repeated DNA these can be purified directly by PFG and used as hybridization probes in partial digests. For other organisms it should be possible to clone these fragments selectively in cosmids or YACs and then either compete away the repeated sequences or subclone single copy DNA. In effect, near the end of a restriction mapping project, it will become efficient to abandon the pure top down approach.

With the bottom up approach, the full impact of highly repeated sequences is unknown but it is likely to complicate matters considerably. Even without this difficulty, cloning biases constitute a major obstacle to completing maps since it becomes progressively more and more difficult to find representatives of those areas that are selected against in whatever procedure was used to make the orginal library. Thus, typically after a certain point in contig building, it becomes necessary to abandon the pure bottom up approach and use other methods to try to bridge between the existing unlinked contigs. These are usually top down strategies. In the future, as new mapping strategies are designed and implemented, it seems inevitable from the results of the early studies described above, that hybrid methods, combining elements of both top down and bottom up will prove to be the most efficient.

What are needed to expedite both existing mapping methods are more efficient ways of focusing on a particular region of a chromosome. One

way that this has been done is to use hybrid cells containing only the desired region, say, of a human chromosome, in an otherwise neglectable rodent background. New methods for introducing selectable markers into desired chromosome locations would really increase the generality of this approach. Microdissection of chromosomes would appear to be an even more attractive alternative for future studies (5). Earlier studies using this approach were plagued by the scarcity of the resulting material. However, appropriate use of PCR should circumvent this difficulty readily.

RESOLVING DISCREPANCIES

Any real scientific effort is compromised by potential experimental and interpretative errors. As large genomic areas are mapped and cloned in independent parallel efforts, the problem of finding and resolving errors looms as quiet a serious obstacle. Both the shear mass of data that goes into a mapping effort, and the error-prone nature of some of that data are likely to cause difficulties. Even in simple mapping efforts, scores of clones, and hundreds of Southern blots are handled. The most likely source of error will probably be interchanges of samples: clones, gel lanes, hybridization mixtures, restriction enzymes, and so on. In practice, even within one coordinated laboratory effort, it is very difficult to discover and correct such errors when so many different samples are involved. A second major source of error is inaccuracies in translating detected DNA fragments into finished physical map. In most current top down or bottom up mapping methods, sizes of fragments are critical input data. Sometimes these can be misjudged because a weak band is missed, a loading artifact alters apparent sizes, or the manual reading of a blot simply fails to include a key data point. The difficulty is the subjective nature of the initial data record, compounded by the inexact algorithms actually used to assemble maps from raw data. The problem is surely even worse when the maps are assembled by hand.

A third potentially serious source of error is unresolved differences in the names and identities of clones and cell lines used in mapping and sequencing efforts. An example of the sort of problem we will face is shown in Fig. 2. Are the differences experimental error, or just sample identities?

What will be needed is much more objective data analysis, and much more robust methods to scan large amounts of data looking for

inconsistancies. These methods inevitably will be very computer intensive, and they will not be simple to implement given the complex nature of gel images, with inherent distortions and the need to integrate such images with documentation and sample descriptions Ultimately one will have to have the capability to reanalyze large amounts of map data without significant human intervention. Only in this way are we likely to be able to distinguish discrepancies that arise from different map construction algorithms from those which arise from accidental or unrealized differences in raw data. A few examples of the kinds of problems that must be faced are as follows:

How can we tell if a clone used in one mapping effort is incorrect and has produced a discordant map?

How can we tell if a cell line used in one mapping effort is either incorrect, or else has a genuinely different map from those used in other efforts?

How can we tell if a key gel image is wrong because two lanes have been interchanged or because the samples or probes as a whole are not what they were thought to be?

The challenge to map makers is to develop the tools needed to address these and similar questions soon, before the amount of potentially discordant data becomes overwhelming.

DEALING WITH POLYMORPHISMS

The genomes of different human beings are not the same. It has been estimated that about one nucleotide position in a thousand is polymorphic to a significant extent in the existing human population. This estimate is probably biased towards unique sequence and coding regions so the true extent of polymorphism may be considerably higher. The polymorphism arises both from true DNA sequence differences and from differential cytosine methylation. It allows powerful diagnostic tests to be developed to resolve paternity and forensic issues. The polymorphism also forms the basis of current human genetic mapping efforts (6,7) as well as the powerful new diagnostic tools available once linked polymorphic DNA markers have been found near genes responsible for human diseases (8). However the extent of human DNA polymorphism also poses a potential

difficulty for human mapping and sequencing efforts.

The fact that mammalian cells are diploid means that the effects of polymorphism are felt even if one choses to work with a single clonal cell line. In general a physical map will reflect the polymorphisms present. For example consider a restriction site present as a heterozygous locus. If one ignored the potential for such heterozygosity, the site would appear to have arisen from a partial digest since it would be cleaved in only half of the DNA molecules. Most of the enzymes used for top down mapping are sensitive to DNA methylation and many methylation sites are only partially modified in typical cell lines. Thus the effects of methylation and true heterozygosity cannot always be readily distinguished. In practice, though, one can circumvent most of the potential ambiguities by recognizing that a restriction map will represent sites potentially cleaved, but not all will necessarily be cleaved in a given cell. This notion can be generalized to cover sets of different cell lines where substantial map differences may occur. In practice, we have found it convenient to make physical maps using a family of 8 to 16 unrelated cell lines. Most regions will be quite similar among the lines. Where differences occur, they actually serve as a fingerprint for the particular region of the chromosome. Analysis of the pattern of differences among the lines is a great asset in determining whether two DNA probes accidentally lie on similar sized unrelated DNA fragments or in fact lie on the same DNA fragment.

Extending this notion permits much more powerful analysis of partial restriction digests because the polymorphic sites help resolve left-right ambiguities in interpreting such digests. However it is important to realize that the resulting restriction maps of polymorphic regions will not necessarily correspond to the true map in any cell line unless it can be shown that the line is homozygous. In practice what can be done is to include several hybrid cell lines among those in the mapping panel. These are usually homozygous for their human chromosome. The hybrid cell is used as the major reference line for the final map. The result is a true homozygous map of that line, but unfortunately it will reflect a rodent DNA methyation pattern, and suffers from the risk that the hybrid cell could easily harbor chromsome rearrangements that would not be tolerated in a human cell.

The above discussion should make it clear that the question "who to

map" is not a serious one. The need to rely on hybrid cells for mapping means that it is unlikely a reference map will derive from a single individual. The need for hybrid cells containing altered chromsomes such as deletions or translocation, makes it even less likely. The human physical map and the ultimate sequence are likely to be a composite of many unrelated individuals. This seems appropriate, actually, in view of the global nature of the project.

FINDING GENES

The major goal of the human genome project is finding the 50,000 to 100,000 estimated human genes. These are the major interest for future biological studies and the major target for initial sequencing efforts. We are interested in making physical maps because these assist the search for genes. The most direct impact is that a physical map, especially an ordered library. allows direct access to DNA molecules that are reasonable candidates to contain a gene of interest. However, in practice there can be a number of serious complications.

The human genetic map, today, has an average resolution of about 10 million base pairs (Mb). In such a region there are likely to be 300 genes. Available genetic information will place a gene of interest within a particular 10 Mb interval; sometimes there will be some hint of where it lies within that interval. If sufficient family material exists, it is then possible to construct a finer genetic map, and perhaps eventually to localize the desired gene to within 1 Mb. For very rare diseases this will not be possible. Even for more common diseases, there will be difficulties in many cases. Human genetic linkage studies are based on meiotic recombination frequencies. These are not uniform throughout the genome and we cannot rule out order of magnitude distortions between the genetic and physcial map in any particular region. Even putting this issue aside, there are other problems in trying to refine genetic mapping even further. It seems reasonable to speculate that much of the observed human meiotic recombination will occur at localized hot spots (9). Between such regions no information will be extractable by pure genetic approaches. Thus the ultimate human genetic map may consist of markers arranged in ordered blocks containing sets of genes, but no genetic order will be observable within each block. We can only guess today at the likely size of such blocks, but they will probably contain a goodly number of genes each.

Genetics must be used to narrow down the location of most genes as much as possible since no other method can deal with a disease allele known only through its phenotype. What does one do, then, once the available genetic resources have been exhausted? The answer really depends on the nature of the disease alleles available. To date all disease genes that have been found by linkage and ultimately cloned have been assisted greatly by the availability of alleles with signficant DNA rearrangements. The utility of such alleles is clear: once one has approached the neighborhood of a gene, such a allele becomes visible as a direct alteration in DNA size. The aerial view provided by physical maps immedately pinpoints the location of the potential disease gene.

In the absence of a DNA rearrangement, the search for a disease gene will be much more tedious. One fortunate phenomenon is that many, perhaps most genes are accompanied by peculiar DNA regions called CpG or HTF (Hpa II tiny fragment) islands (10). These regions, typically 1 kb in length and located 5' to a gene, show no CpG methylation; they are G-C rich, and are particularly rich in CpG. This leads to the observation that restriction enzymes that cleave at unmethylated CpG sequences, cut almost exclusively in such CpG islands. As a result, the sites for such enzymes are clustered in the genome. Each cluster is an unmistakeable signature for the presence of a gene. Thus restriction maps, which could have just represented arbitrary sequence locations, in fact tend to be maps of gene locations. Once genetics has led to the focus on a region, the restriction map will readily provide candidate sequences for expressed genes. These can be used to study interspecies conservation and direct tissue-specific patterns of expression to try to exploit whatever physiological hints exist about a particular disease to infer which of the many genes in a region are attractive candidates for the disease gene.

In the most difficult cases, a region may present itself as an indivisible cluster of many genes, all with no apparent rearrangements in disease, and all with no compelling patterns of tissue specific expression. Unless one can demonstrate altered levels of expression in a diseased individual, the task of determining which of the genes in the region is responsible for the disease may have to rest on extensive DNA sequencing. In a typical 1IMb DNA region there will be 1000 polymorphic single base loci. Only one of these need be the disease locus. It will show an exact correlation with the disease phenotype. The problem is how to find it. By concentration on cDNAs one may be able to narrow the search by an order

of magnitude. However the amount of sequence that will have to be determined is still considerable. Simple calculations suggest that a set of five unrelated disease-carrying individuals and five normal controls would probably suffice. The task will be to obtain and sequence the DNA from these individuals. PCR promises to simplify this task quite considerably. It will greatly accelerate the preparation of the desired DNA samples since one reference DNA sequence of the region will indicate the optimal set of primers needed to determine the DNA sequence from all of the individuals in the study.

INFERRING GENE FUNCTION

Eventually we will know the DNA sequence of tens of thousands of previously uncharted genes. Except for those that can readily be localized to a few thousand genetic diseases, or those that are close relatives of genes already studied in other organisms, we will have few clues about the function of these genes. The likelihood that many will function only in the central nervous system is a further obstacle.

The first step in studying each unknown gene will surely be a comparative search among genes with previously known function. Unfortunately the three dimensional structure of proteins is better conserved than their sequence. The limited amount of available tertiary structural information will surely be frustrating and one must hope that by the time the human DNA sequence is available, methods for directly interpreting it will allow the inclusion of at least some inferences about plausible three dimensional structures. The likelihood that this will occur probably depends heavily on the rate of experimental structural determination during the next decade since inferences from known structures appear to be the most powerful tool currently available in sharpening our currently very limited structure prediction abilities.

Even with a known three dimensional structure, it is very difficult to make reliable estimates of protein function. Experimental approaches will be needed to go much further than we can at present. Finding and interfering with comparable genes in model organisms such as mice and drosophila is probably the best strategy we have at present. It remains to be seen how general this approach will be, especially for highly specialized functions such as may exist in the human central nervous system.

We should, however, not be too pessimistic. Protein structures have arisen by evolution. They are related and many related structures have related functions. Motifs, structural elements of proteins, have been shuffled around as an efficient way of rapidly creating new, complex functions. The significant structural and energetic constraints imposed by the necessity of proteins to fold and achieve stable structures is apparently responsible for the recurrence of a relatively limited set of motifs. This pattern in protein structures, once we understand it better, may greatly simplify the taks of classifying and assigning putative functions to the deluge of new genes. The analogy in Fig. 3 is a fairly obvious clue to the way in which the same elements used in different combinations can yield different, but related, functions.

DISTRIBUTION OF DATA AND SAMPLES

The human genome project is going to generate enormous amounts of data and enormous numbers of samples. To manage this material and make it optimally accessible and useable to the broad community of biological scientists is going to require significant changes in our style of work. First it seems necessary that all this material reach the public domain rapidly and freely and that a frequent accounting of what is available be made. Whether all the material is stored in a single physical location, one database, one sample archive, seems unimportant. However the prospective user should not have to search numerous partially redundant catalogues: the storage system must have the appearance and the efficiency of a single site even if it is not this way in practice.

The details of how sequence and map data should be stored, integrated and distributed to the public are not agreed upon yet. We are even further from knowing how to proceed with biological samples on a scale that dwarfs the sum of all current clone and cell line repositories. However it seems clear that the data can only be handled in a fully computerized way, and the samples in a fully automated way. Developing the ability to do this is a major task for the genome project and it is likely to consume a fair share of the available resources. PCR many have eliminated some of the difficulties in handling large clone banks. Once the DNA sequence of a reference genome is known, any desired short stretch of that sequence could be easily compared by PCR. Thus it may not be necessary to distribute the clones, just the cell lines from which the

DNA sequence was derived. For this reason the stability of the cell lines used as sources for large scale sequencing efforts may be a very important consideration.

If the human genome project were carried out at a single location one might be able to standardize data base and sample handling protocols at an early stage. This would simplify many of the problems described above. However given the dispersed nature of the ongoing project, and the numerous scientific and practical reasons for maintaining this nature, there are unlikely to be uniform choices of hardware, software, cell lines, restriction enzymes, nomenclature, and so on. The best we can hope for is that sophisticated software will be developed that will shield the user from all but essential differences resulting from the idiosyncratic styles of individual genome efforts. If this single goal can be achieved it will have an impact far beyond the genome project, even far beyond biology since already a vast array of computer software and hardware differences complicates all data-intensive pursuits.

Acknowledegments: This work was supported by grants from NIH (GM 14825), NCI (CA 39782), and DOE (DE-FG02-87ER-GD852).

FIGURES

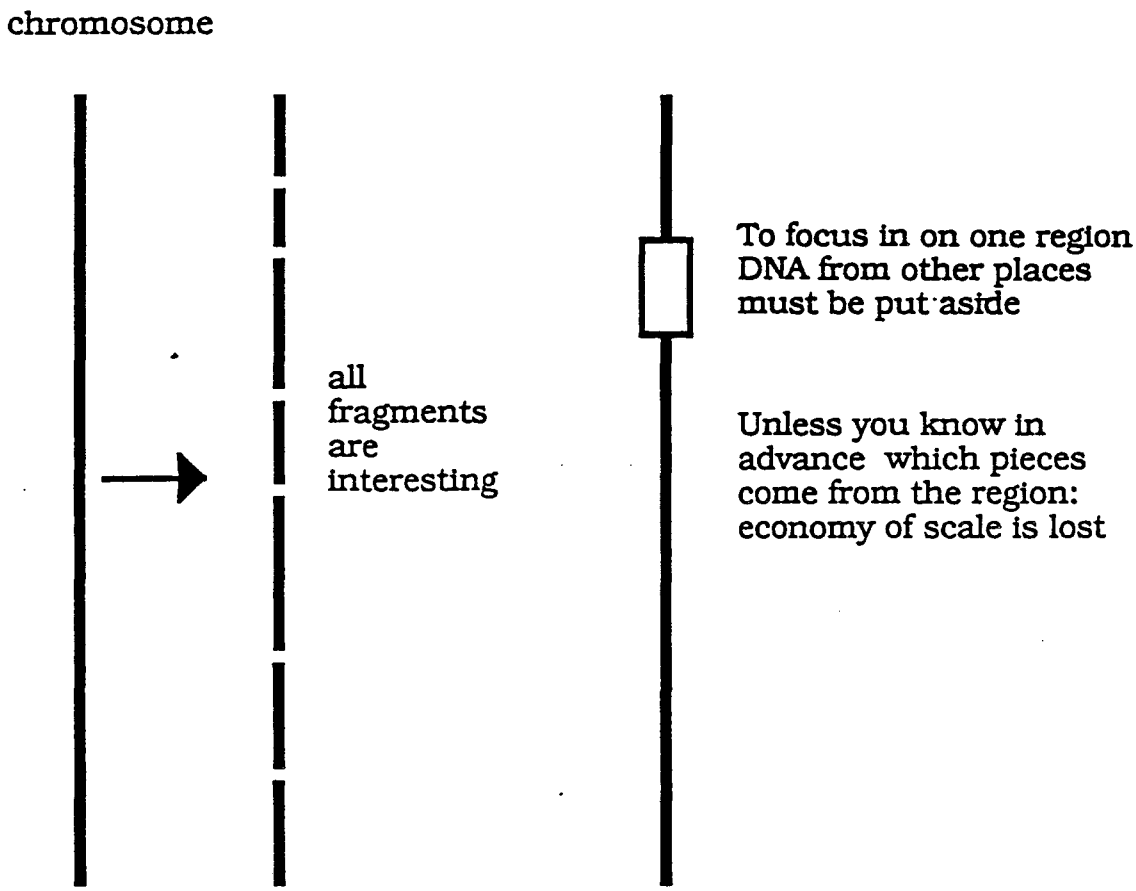
Figure 1. Finishing a map is difficult.

Figure 2. Resolving discrepancies: shown are three schematic Southern blots. How can we determine if the differences obtained by the laboratories are significant or just represent problems in nomenclature and electrophoretic running conditions?

Figure 3. Structural motifs can be recombined to yield different but related functions, here in mechanical conveyances but the same notion should be applicable in proteins.

REFERENCES

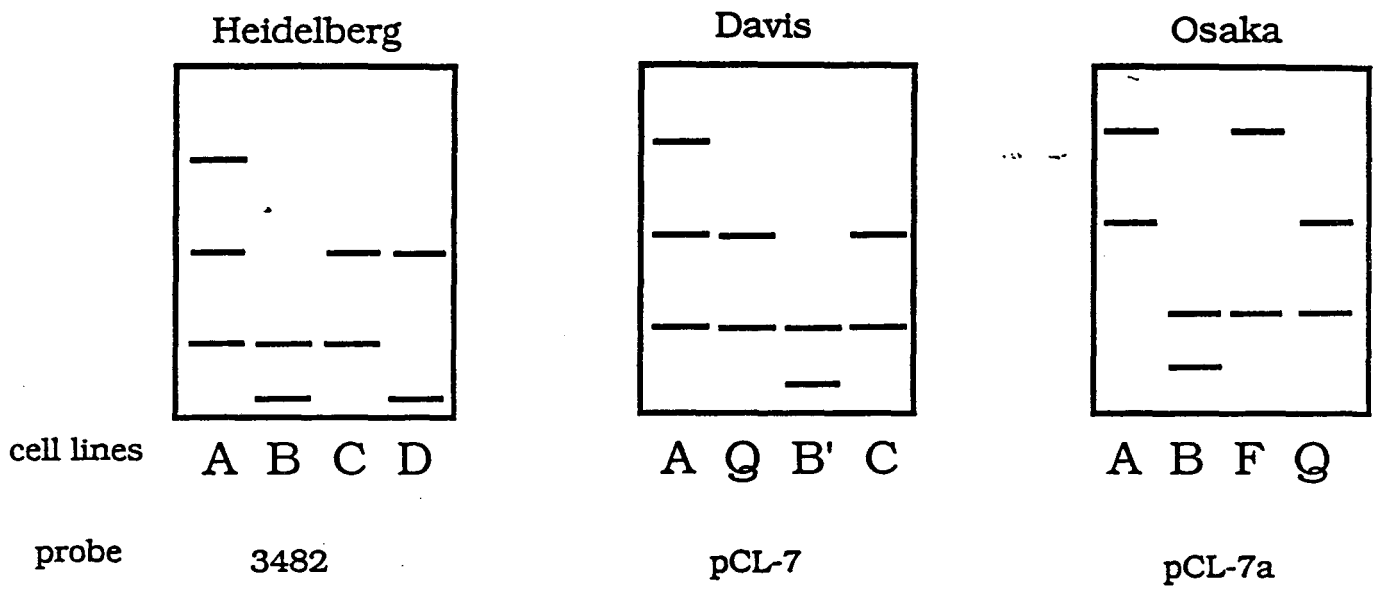
1. Kohara, Y., Akiyama, K., and Isono, K. (1987) *Cell* **50**, 495-508
2. Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., and Frank, T. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7826-7830
3. Coulson, A., Sulston, J., Brenner, S., and Karn, H. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7821-7825
4. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., and Horn, G.T. (1988) *Science* **239**, 487-491
5. Lüdecke, H.-J., Senger, G., Claussen, U., and Horsthemke, B. (1989) *Nature* **338**, 348-350
6. Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S., Botstein, D., Akots, G., Rediker, K.S., Gravius, T., Brown, V.A., Rising, M.B., Parker, C., Powers, J.A. Watt, D.E., Kauffman, E.R., Bricker, A., Phipps, P., Muller-Kahle, H., Fulton, T.R., Ng, S., Schumm, J.W., Braman, J.C., Knowlton, R.G., Barker, D.F., Crooks, S.M., Lincoln, S.E., Daly, M.J., and Abrahamson, J. (1986) *Cell* **51**, 319-337
7. White, R., Leppert, M., O'Connell, P., Nakamura, Y., Julier, C., Woodward, S., Silva, A., Wolff, R., Lathrop, M., and Lalouel, J.M. (1986) *Cold Spring Harbor Symp. Quant. Biol.* **51**, 29-38
8. Royle, N.J., Clarkson, R.E., Wong, Z., and Jeffreys, A.J. (1988) *Genomics* **3**, 352-360
9. Steinmetz, M., Uematsu, Y., and Lindahl, K.F. (1987) *Trends in Genetics* **3**, 7-10
10. Antequara, F., and Bird, A.P. (1988) *EMBO J.* **7**, 2295-2299



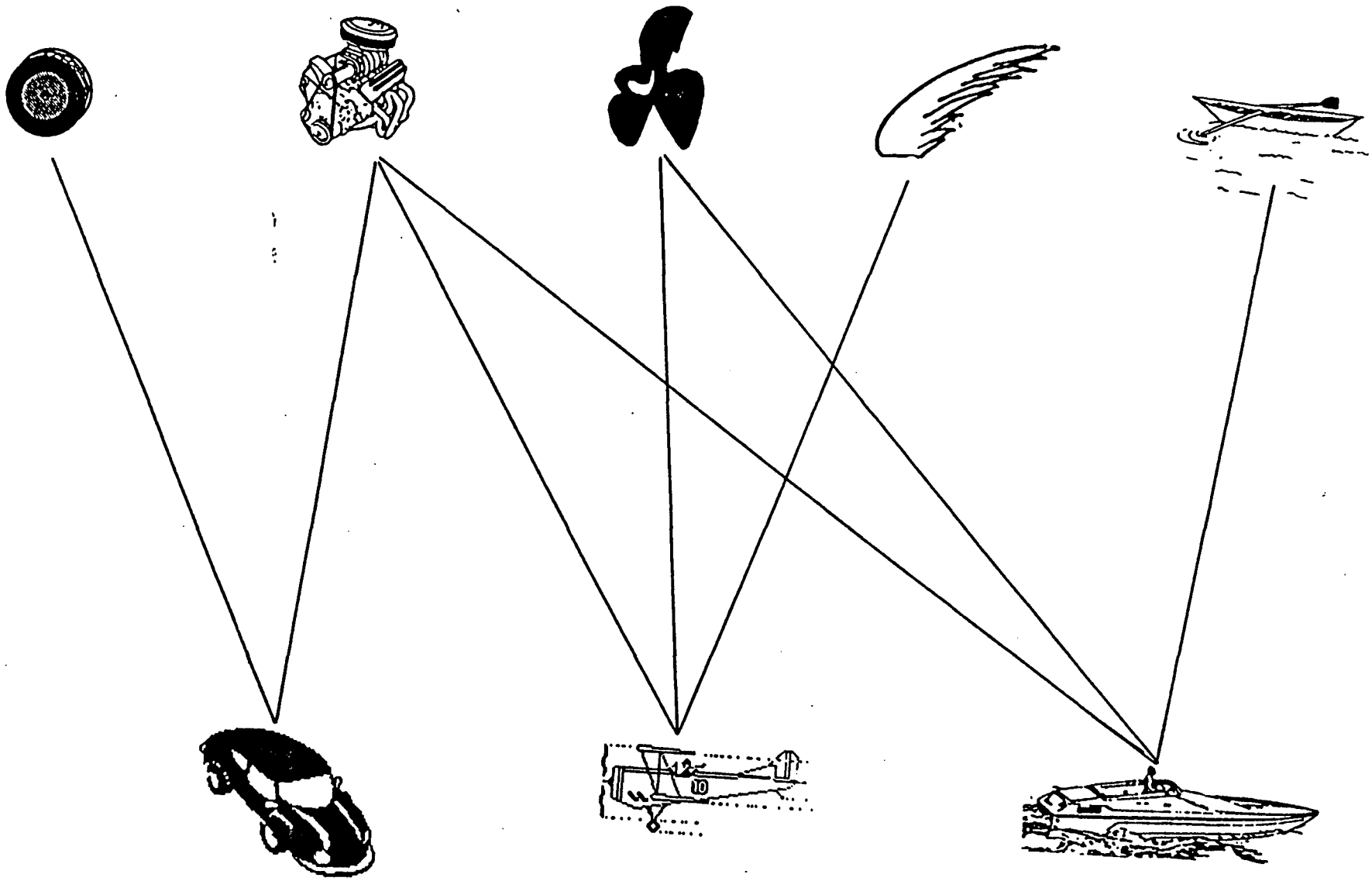
Economy of Scale

Finishing a map is difficult

Resolving discrepancies?



Resolving discrepancies: shown are three schematic Southern blots. How can we determine if the differences obtained by the laboratories are significant or just represent problems in nomenclature and electrophoretic running conditions.



Structural motifs can be recombined to yield different but related functions, here in mechanized conveyances but the same notion should be applicable in proteins.

LAWRENCE BERKELEY LABORATORY
TECHNICAL INFORMATION DEPARTMENT
1 CYCLOTRON ROAD
BERKELEY, CALIFORNIA 94720