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.

**SIMULATIONS OF ORDERING AND SEQUENCE RECONSTRUCTION
OF RANDOM DNA CLONES HYBRIDIZED
WITH A SMALL NUMBER OF OLIGOMERIC PROBES**

ANL/BIM/CP--77874

DE93 004177

IVAN LABAT and RADOJE DRMANAC

Integral Genetics Group

Biological and Medical Research Division

Argonne National Laboratory, Argonne, Illinois 60439, USA

Received by OSTI

DEC 09 1992

ABSTRACT

The sequencing by hybridization (SBH) method has been developed for assaying millions of 0.5- to 2-kb-long clones. This opens up an efficient way for defining the order of short clones and creating a physical map at 100-bp resolution. Moreover, complete sequences can be obtained using a modest number (about 3000) of probes if hybridization and gel sequence data from overlapped or similar sequences are used. In light of these possibilities, various heuristic algorithms have been developed and tested

The submitted manuscript has been authored by a contractor of the U. S. Government under contract No. W-31-109-ENG-38. Accordingly, the U. S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U. S. Government purposes.

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

in simulation experiments. This approach can influence the interpretation of the intuitively obvious term, "known sequence."

Introduction

In our approach, knowledge of the intensity of the hybridization signals from a given set of oligonucleotide probes with a given certain piece of DNA is used as a source of information for mapping and sequencing purposes. Hybridization experiments using different sets of probes can provide sufficient information for recognizing sequence characteristics at different levels of resolution. Starting from a small set of probes (200 6- to 8-mers), one can obtain sufficient information for clone ordering and structural maps. Upgrading the set to about 3000 6-to 8-mers, and using complementary data from similar sequences and from the other, overlapped clones of the same sequence, a complete sequence (i.e., the ultimate map) can be obtained.^{1,2} We have developed and tested algorithms for clone ordering and used information from similar sequences to completely reconstruct each sequence from less redundant hybridization data.

An Algorithm for Defining the Order of Random M13 Clones

The basic concept of our approach to clone ordering is to a degree similar to the idea of defining informational fragments (IFs) in complete sequence reconstruction.³ IFs are defined by the beginnings and ends of overlapped clones. Therefore, one can define

the overlapped (and nonoverlapped) portions of individual clones by determining the intersections and differences of sets of probes that hybridize to the clones of interest. Using those, the order can be obtained. The scheme of the process is shown in Fig. 1.

The process is as follows. Starting with any clone A, all of the clones containing a certain percentage (usually 60%) of probes shared with clone A are selected and arrayed by decreasing number of shared probes. Starting from the first clone in the array i.e., the one most similar to clone A, a subset (B) of clones from the array is selected that satisfies the following rules:

- (i) There is a subset of probes (r) shared among clone A and the clones in the B list.
- (ii) The remaining probes hybridizing with clones in B can be separated in two groups, r_l and r_d , such that:
 - (iia) each of probes from r_l hybridizing to a clone from B is present in all succeeding clones in the B list,
 - (iib) Each of the probes from r_d hybridizing to a clone from B is present in all succeeding clones in the B list.

The order of the clones in B is determined by comparing the subsets r_l and r_d . From figure 1, it is obvious that probes r_d present in the succeeding clones should not be

present in the previous one(s) and probes *rl* are not present in the succeeding clone(s) in the array.

When the primary contig is formed, subset *rn* is defined from the subset *rd*. Now, instead of selecting clones sharing 60% of the probes, clones containing all probes from subset *rn* are selected too. Due to the repetitiveness of the genomic sequence, a number of primary contigs can be formed starting from one clone. These branches are on Fig. 1., represented by dashed and empty bars. Some of these (represented by the branch of empty bars) are not physically overlapped with clone A, although they contain 60% or more probes shared with clone A. Two of these branches have a physical overlap (5' and 3' branches). This causes an ambiguity in determining the actual (physical) position of the starting clone. Due to this, the starting clone is not included in any of the branches, except in the case when a single branch can be defined.

The clone-ordering algorithm has been successfully tested on a randomly created sequence of cosmid size (40,000 bp), which was randomly "subcloned" in 400 clones of 1000 bp, achieving tenfold coverage. For the test, we used 200 randomly chosen 6-mer probes. Thus, each probe had a 25% chance to be present in each of the clones (i.e., each clone, on average, contained about 50 probes). The unique order was obtained in 59 cycles.

For ordering clones from considerably longer sequences (in present experiments we are using a 350,000-bp sequence represented by about 3500 clones in either orientation), the problem of branching is more serious. We are developing another

process in which branches will be connected unambiguously. This process is repeated until all clones are used.

In conclusion, in an ideal case, N -fold coverage defines $2 \times N-1$ IFs on one clone. Therefore, tenfold coverage of a certain sequence with clones 1000 bp long defines IFs of about 50-100 bp, i.e., about 19 informational fragments on each clone. The probability of unique identification of that particular clone is about 10^{17} times increased when considering it as an ordered set of 19 informational fragments rather than as a whole clone. This holds for the assumption of random sequence.

Structural Maps

The physical map of a sequence contains information on the position of and distances between specific sequence markers. It does not provide, however, any information on the sequence segments between the markers. The application of partial sequencing by hybridization,^{2,4} when hybridized probes read over 10% of base pairs, provides information on the genome structure, that is, the order and distance relationships of various sequence elements of interest.²

For our experiments, we compiled eight of the longest human genomic entries from GenBank, thus simulating the real, human sequence for the length of 350 kb. On this sequence, we defined the most frequently occurring probes. Our assumption was that the most frequently occurring probes will be those that are the most informative for mapping and partial sequencing. Due to local variations of G+C content in the sequence

(the coding regions of vertebrates have 10-20% higher G+C content than noncoding regions),⁵ we selected the most frequent probes of three categories, postulating that probes of different C+G content would be preferentially present in different regions of the genomic sequence. Our choice of the categories of probes was:

- a) 60 8-mers consisting of A or T only (W)
- b) 66 7-mers containing of one or two C or G (AW)
- c) 74 7-mers containing 3 or more C or G (CG).

An additional caveat for the choice of probes was that each probe had to have its reverse complement in the set. This is introduced for simple reason that in the M13 cloning process, either strand of a given piece of DNA can be inserted into the vector. So, instead of 60 W 8-mers, it is better to consider them as 30 pairs of W 8-mers.

The distribution of this set of probes on the sequence is shown in Fig. 2.a. The sequence is divided into 500-bp windows. White dots on one row represent occurrences of a given probe at least once in a given window. From the figure, two things are apparent. First, due to the nonrandomness of genomic sequences, the overall distribution of a certain set of probes is not even. This is particularly obvious for the set of probes with high GC content (c, Fig. 2.a). Second, the different sets of probes are of different importance for obtaining a structural map of the sequence, although the overall frequencies of occurrences of probes from all sets in the genomic sequence are similar.

By comparing the sequences in windows with the most frequent occurrences of (c) probes with the sequence description lines of GenBank entries, it was shown that these

sequences coincide mainly with the positions of Alu repetitive elements. For comparison, we performed the same operation on a nongenomic 350-kb sequence (Fig. 2.b). To generate a nongenomic sequence with the same base composition (and distribution), we simply reversed the orientation, i.e., defined the 3'-5' strand of our original sequence. From the figure, it is obvious that with the CG probes, it is impossible to define any structure of the sequence above the level of a random smear, which is not the case with the W and AW probes. A comparison of the distribution of the CG set of probes on both sequence orientations is given in Fig. 3 a,b. In original sequence orientation, different probes of this subset are present in various windows having the high content of CG probes, that is, in different Alu sequences. This is result of the fact that the average similarity between a member of the Alu family and the Alu consensus sequence is about 80%.⁶

Use of Similar Sequences in Complete Sequence Reconstruction

The major problem in the reconstruction of a sequence from its constituent oligonucleotides is the occurrence of branching points. If L is the length of oligonucleotides considered, branching points are defined by repeats of sequences $L-1$ in length. These repeats define ends of sequence subfragments.³

The only possibility of overcoming this problem is in reconstructing the sequence from a list of oligonucleotides lacking all but one repeat. We are handling branching points in two ways. The first is by using the information on the oligonucleotide content

of the overlapped sequences. In reconstructing an overlapped sequence that does not contain one of the $L-1$ sequences present in the original one, that particular branching point does not exist. The use of this approach is described elsewhere.⁷ The second possibility is the use of similar sequences in which either of $L-1$ sequences is altered.^{1,2}

To test the discriminating power of the second method, we have performed experiment with two 90%-similar sequences, each 150 bp long. Using a complete set of 6-mer oligonucleotides, these two sequences were reconstructed to a point where sequence subfragments could be connected in 30 and 40 ways, respectively. Results of about 1200 pairwise comparisons of these solutions are shown in Fig. 4. As the graph shows, only one pair of possible solutions has the original similarity of 90%. The sequences in this pair are the correct ones. All other possible pairs show at least 20% lesser similarity. There is, however, a certain possibility of obtaining a pair of solutions showing similarity higher than the original one, i.e., that the pair of solutions showing the highest similarity is not the pair of correct sequences. This is especially the case when short segments of a long sequence are reconstructed.

To check this, we performed experiments similar to the one described. In these experiments, sequences were elongated to an assigned length (usually 20 bp), and possible solutions were checked for similarity. In a series of reconstruction experiments, we found that in about 12% of the cases, the similarity of an incorrect pair of sequences 20 bp long can be up to 20% higher than the original, correct pair of sequences (Fig. 5). This implies that the similarity threshold for correct sequences in the reconstruction algorithm

has to be lowered 20% from the original similarity for up to 20-bp elongations, to ensure that the correct solutions are not omitted.

In another simulation experiment, we attempted to reconstruct a 500-bp sequence using a complete set of 6-mers as probes and combining information from overlapped and similar sequences. For this, two random, 90%-similar sequences, each divided into three 500-bp clones overlapped by 50% were simulated. This coverage defined informational fragments of 125-250 bp. The setup is shown in Fig. 6. Reconstruction was performed on clone c2 from G1. As a result, clone c2 was reconstructed to a level of five sequence subfragments, defining two possible solutions. In situations like this, only one additional bit of information would have been sufficient for complete, unambiguous sequence reconstruction. This is far less than in the reconstruction a single 500-bp sequence using 6-mers, which would result in forming several hundred sequence subfragments.³ The large quantity of additional information needed would make a complete reconstruction very inefficient if the sole 500-bp fragment is used.

The influence of a level of similarity on the efficiency of sequence reconstruction was also measured. We used random sequences 300 bp long and a complete set of 6-mers as probes. The decrease in the number of possible solutions to sequence reconstruction using pairs of similar sequences was measured. The results are shown in Fig. 7. Sequences 70-90% similar gave the best results. Using the setup as described, four similar genomes would suffice for unambiguous sequence reconstruction of each. Genomes similar at a level above or below the range of 70-90% are less useful for this

purpose. A higher number of possible solutions implies the need for a larger number of genomes for complete reconstruction.

It is important to note that the use of several genomes would not increase the amount of work needed for reconstruction. On the contrary, reconstruction of a single genome sequence requires more work, in sense of using longer probes (i.e., a larger number of probes). In the approach described, the decrease in the number of probes (probe length) several times outweighs the increase in the number of clones needed. The number of required hybridization experiments is:

$$C \times 4^L,$$

where L is length of the probes used for unique reconstruction and C is the number of clones of adequate length. For example, complete and unambiguous reconstruction of a single sequence divided in 300-bp IFs implies the use of a complete set of 9-mer probes.³

The number of experiments to be performed is: $C \times 4^9 / (C \times 4 \times 4^6) = 16$,

i.e., 16 times larger than the number of experiments needed for sequencing of four genomes 70-90%-similar using 6-mers.

Sources of similar sequences are inter- and intragenomic. Many genomes show a similarity with the human genome within the optimal range. In certain coding elements, even genomes as evolutionarily distant as cow and human show the similarity level required.⁸ Some, on the other hand, are too similar, as is case with the 1-5% difference between human and some other primate genomes.⁹

Members of gene families are good examples for intragenomic similar sequences that can be used in sequence reconstruction. Alpha and beta globins,¹⁰ alpha and beta interferons¹¹ showing over 50% homology, as well as genes diverged more recently, such as beta major and beta minor mouse globin genes (over 90%),¹² have extremely useful intragenomic similarities. Also, known consensus sequences of certain genomic elements (e.g. Alu, LINE1) can be used as sources of information needed for reconstruction. Members of the mentioned sequence families are 80%⁶ and 90%¹³ similar to their consensus sequences, respectively.

Discussion

The described methods have the potential to provide a high resolution in mapping of DNA sequences. The resolution itself depends on two factors. The first is the way the genomic library is made in terms of clone length and the multiplicity of coverage of the sequence. For example, ten times coverage by 1- to 2-kb clones defines resolution at the level of 50-100 bp for fragments of DNA (IFs) defined by beginnings and ends of overlapped clones. The second factor is the length and sequence of the applied oligonucleotide probes. Information value of the probes depends on the frequency of appearance of a given probe (or group of probes) in the sequence of a fragment of given length. More precisely, it depends on different frequencies of probe appearance in a given genomic sequence and in a random sequence. Existing methods for determining

the physical map via clone ordering¹⁴⁻¹⁶ use the different lengths of genomic fragments obtained by restriction enzyme digestions.

Applying the calculations of Lander and Waterman,¹⁷ using these methods to obtain a physical map of a resolution of 100 bp from 0.5- to 2-kb-long clones would require implementation of up to 80 different 6-cutter restriction enzymes. This makes these mentioned approaches impractical for the ordering of short fragments. Other proposed methods use oligonucleotide hybridization fingerprinting techniques,¹⁸ similar to the approach described here. However, in the existing fingerprinting approach, the use of likelihood ratios and minimal spanning tree algorithms is assumed in order to obtain the order of clones. The use of a minimal spanning tree does not require information about which probes differ between pairs of clones in an array of overlapped clones; only information about their number is required, thus reducing the overall information value of the system.

In our model system for determining the order of short clones overlapped so that the sequence from which they are derived is covered ten times, it is sufficient to apply a set of 200 6-mer probes. The expected probability of occurrence of a certain 6-mer oligonucleotide in a clone 1000 bp long is about 25%. Therefore, a 100-bp informational fragment should be represented by a unique group of five probes from a set of 200, provided that the probes are uniformly distributed along the sequence.

In conclusion, for purposes of clone ordering, the aim is to obtain a sufficient number of uniquely defined informational fragments. For this, the coverage of clones with probes should be uniform, and probes should be independent of each other in their

occurrences in the sequence. On the other hand, defining structural maps poses a completely opposite requirement. The uniform coverage of any sequence with a probes from a certain set would not provide information on sequence structure, regardless of the number and positions of repetitive sequences.

Different sequence characteristics, such as AT-rich non-coding elements, CG-rich Alu repeats, etc., can be used as sources for the definition of sets of probes for structural mapping. The choice of a set of probes satisfying both the clone ordering and structural mapping requests would, therefore, probably consist of a number of subsets, each recognizing one of the major sequence determinants. Improvement of the clone ordering algorithm would include the use of data on the intensities of the hybridization signals, rather than an all or none approach. Lower intensities of hybridization signals can come from a number of mismatched hybrids, rather than one full match. Two clones with the same quality of signals obtained with a given set of probes are more likely to be overlapped than two clones giving a different quality of signals with the same set.

The similar-sequences approach drastically reduces the number of probes needed for complete sequence reconstruction. Moreover, the beginnings and ends of two libraries from similar sequences covering the same region define shorter informational fragments on both, thus further reducing the requirements for a number of probes to be applied (Fig. 6). Using the described method, the complete sequence of a human genome, along with 3-4 other, similar genomes could be obtained applying a complete set of 6-mer probes. For the human genome alone, the same task would require a complete set of 9-mers.

In microbial genomics, one can envision a number of strains of microorganisms helping one another to be sequenced. In higher organism sequencing, lacking mutant strains, similar sequences from genomes of different species or similar parts from one genome (repetitive elements, gene families, etc.) have to be used. As mentioned before, the optimal similarity for these purposes lies in the range of 70-90%. Once one of the sequences is completely determined (using other genomes of appropriate similarity), it will be used as a template for the unambiguous reconstruction of the sequence of high similarity in other, closely related genomes.

Acknowledgments

We thank David E. Nadziejka for technical editorial assistance. This work was supported by the U.S. Department of Energy, Office of Health and Environmental Research, under contract No. W-31-109-Eng-38.

References

1. R. Drmanac, *Genome Mapping and Sequencing*, Cold Spring Harbor Laboratory Press, 1992, 318.
2. R. Drmanac and R. Crkvenjakov, *Int. J. Genome Res.* **1** (1992) 59-79.
3. R. Drmanac, I. Labat, I. Brukner and R. Crkvenjakov, *Genomics* **4** (1989) 114-128.
4. R. Drmanac, G. Lennon, S. Drmanac, I. Labat, R. Crkvenjakov and H. Lehrach, *Electrophoreses, Supercomputing and the Human Genome*, eds. C. R. Cantor and H. A. Lim (World Scientific, Singapore, 1991), pp 60-74.
5. G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunierrotival and F. Rodier, *Science* **228** (1985) 953-958.
6. Y. Kariya, K. Kato, Y. Hayachizaki, S. Hirzeno, S. Tarui and K. Matsubara, *Gene* **53** (1987) 1-10.
7. R. Drmanac, I. Labat and R. Crkvenjakov, *J. Biomol. Structure and Dynamics* **5** (1991) 1085-1102.
8. U. Galili and K. Swanson, *Proc. Natl. Acad. Sci. USA* **88** (1991) 7401-7404.
9. M. M. Miyamoto, B. F. Koop, J. L. Slightom, M. Goodman and M. R. Tennant, *Proc. Natl. Acad. Sci. USA* **85** (1988) 7627-7631.
10. Y. Nishioka and F. Leder, *Cell* **18** (1979) 875-882.
11. T. Taniguchi, N. Mantei, M. Shwarzstein, S. Nagata, M. Muramatsu and C. Weissmann, *Nature* **285** (1980) 547-549.

12. D. A. Konkel, J. V. Maizel, Jr. and P. Leder, *Cell* **18** (1979) 865-873.
13. A. F. Scott, B. J. Schmeckeper, M. Abdelrazik, C. T. Comey, B. O'Hara, J. P. Rossiter, T. Cooley, P. Heath, K. D. Smith and L. Margolet, *Genomics* **1** (1987) 113-125.
14. M. V. Olson, J. E. Dutchik, M. Y. Graham, G. M. Brodeur, C. Helms, M. Frank, M. MacCollin, R. Scheinman, and T. Frank, *Proc. Natl. Acad. Sci. USA* **83** (1986) 7826-7830.
15. A. Coulson, J. Sulston, S. Brenner and J. Karn, *Proc. Natl. Acad. Sci. USA* **83** (1986) 7821-7825.
16. Y. Kohara, K. Akiyama and K. Isono, *Cell*, **50** (1987) 495-508.
17. E. S. Lander and M. S. Waterman, *Genomics* **2** (1988) 231-239.
18. H. Lehrach, R. Drmanac, J. Hoheisel, Z. Larin, G. Lennon, A. P. Monaco, D. Nizetic, G. Zehetner, and A. Poustka, in *Genetic and Physical Mapping*, eds. K. E. Davies and S. M. Tilghman (Cold Spring Harbor Laboratory Press, 1991), 39-81.

Figure Legends

- Fig. 1. Clone-ordering algorithm scheme. Clones (B, solid bars) contain 60% shared probes with the clone A and satisfy the ordering rules (see text). Dashed and empty bars: ordered clones in other branches satisfying the requirement for 60%+ shared probes with clone A; rl, r, and rd: three groups of probes hybridizing to different regions on the clones in B; m: newly defined set of probes from rd, used as a query subset for the next primary contig branch.
- Fig. 2. Structural map of 350-kb human genomic sequence made by merging of eight largest GenBank entries. a, b, and c represent three sets of probes (W, AW, and CG, respectively). a) Structural map of 350-kb sequence; b) Structural map of reverse sequence.
- Fig. 3. Distribution of CG probes on 500-bp windows on a) the 350-kb sequence and b) on the reverse sequence.
- Fig. 4. Similarity among the pairs of possible solutions for two 90%-similar random sequences. Both sequences are 150 bp long.
- Fig. 5. Omission of correct sequence as a function of similarity threshold.
- Fig. 6. Scheme of reconstruction of 500-bp sequence using data from overlapped clones and clones from similar sequence. Sequences G1 and G2 are divided into 500-bp clones (c1-c3), overlapped by 50%. Experiments are performed on clone c2 from G1 using a complete set of 6-mer probes. As

a result, clone c2 is reconstructed to a level where one of two (1-2-3-4-5 or 1-4-3-2-5) combinations of sequence subfragments is the correct one.

Fig. 7. Number of possible solutions of sequence reconstruction expressed as a percentage, as a function of the similarity of the sequences used.

a.

b.

Fig 3K

figo

END

**DATE
FILMED**

6 / 17 / 93