Volume 256, number 1,2, 118–122 FEB 07689 October 1989

An oligonucleotide hybridization approach to DNA sequencing

K.R. Khrapko, Yu P. Lysov, A.A. Khorlyn, V.V. Shick, V.L. Florentiev and A.D. Mirzabekov

V.A. Engelhardt Institute of Molecular Biology, Academy of Sciences of the USSR, Vavilov Str., 32, 117984 Moscow V-334, USSR

Received 2 August 1989

We have proposed a DNA sequencing method based on hybridization of a DNA fragment to be sequenced with the complete set of fixed-length oligonucleotides (e.g., $4^8 = 65\,536$ possible 8-mers) immobilized individually as dots of a 2-D matrix [(1989) Dokl. Akad. Nauk SSSR 303, 1508–1511]. It was shown that the list of hybridizing octanucleotides is sufficient for the computer-assisted reconstruction of the structures for 80% of random-sequence fragments up to 200 bases long, based on the analysis of the octanucleotide overlapping. Here a refinement of the method and some experimental data are presented. We have performed hybridizations with oligonucleotides immobilized on a glass plate, and obtained their dissociation curves down to heptanucleotides. Other approaches, e.g. an additional hybridization of short oligonucleotides which continuously extend duplexes formed between the fragment and immobilized oligonucleotides, should considerably increase either the probability of unambiguous reconstruction, or the length of reconstructed sequences, or decrease the size of immobilized oligonucleotides.

DNA sequencing; Oligonucleotide hybridization, immobilized

1. INTRODUCTION

The projects of mapping and sequencing the entire human and other species genomes are in progress. The major part of these expensive and laborious programs will deal with direct DNA sequencing by the methods of Sanger and Maxam-Gilbert or their variants. Both methods involve a number of rather complicated manipulations including preparation of samples, loading them onto a gel and reading the sequences after the separation of DNA fragments. A considerable reduction in the project costs could be achieved by the automation of existing sequencing methods or by the development of principally novel, simpler ones.

We have recently proposed, in a theoretical paper [1], a novel DNA sequencing method based on hybridization of a DNA fragment of interest to every oligonucleotide of a complete set (i.e. all possible sequences of given length). The suggestion was to immobilize each of the oligonucleotides at an individual dot of a 2-D matrix, thus allowing all the hybridizations to be processed in parallel, with the DNA fragment serving as a probe. Computer-assisted alignment of hybridized oligonucleotides sequences based on the analysis of their overlapping enables one to reconstruct the fragment sequence. However, the occurrence of repeats leads to branching in the process of alignment thus limiting the length of conseqs – the stretches of sequence aligned unambiguously. With the help of specially developed software it was shown that hybridization with the complete set of octanucleotides allowed to read about 80% of randomly chosen fragments 200 bases long.

Several experimental strategies could increase the sequencing efficiency or the length of the sequenced fragments. It was proposed to use as a second step a 'continuous stacking' hybridization of very short oligonucleotides to decrease the uncertainty at branching points and thus increase the length of conseqs. It was also proposed to cut the fragment randomly into short (15–20 bases) sub-

Correspondence address: A.D. Mirzabekov, V.A. Engelhardt Institute of Molecular Biology, Academy of Sciences of the USSR, Vavilov Str. 32, 117984 Moscow V-334, USSR

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/89/$3.50 © 1989 Federation of European Biochemical Societies

fragments in order to disrupt secondary structures that could interfere with matrix hybridization [1].

A method of sequencing by hybridization was independently proposed by Drmanac et al. [2], their approach differing from ours in two main aspects. First, Drmanac et al. have suggested to immobilize DNA fragments, rather than oligonucleotides and to perform serial hybridizations with oligonucleotides of the complete set as probes. Second, it was proposed to align relatively short 'sequence subfragments' by hybridizations with many highly overlapping fragments, rather than to increase the efficiency of each fragment sequencing.

2. MATERIALS AND METHODS

The oligonucleotides were immobilized on a glass plate covered with a 10 μm layer of activated polyacrylamide (the details will be published elsewhere) as 2 mm dots. The hybridizations were performed at 2°C in 2 μl drops of 1 M NaCl, 0.1 M sodium citrate, about 50000 cpm of ^{32}P-labelled probe and, for continuous stacking hybridizations only, an excess of adjacently hybridizing non-labelled oligonucleotide. The washes were carried out in a large volume of the same buffer plus 0.05% SDS at stepwise increased temperatures, each wash being 5 min long. The hybridization signals were measured between the washes at 0°C by a radioactivity monitor through a 5 mm aperture in a lead screen.

3. RESULTS AND DISCUSSION

3.1. *Biochemical procedures*

A simplified scheme illustrating the sequencing approach is presented in fig.1. The first step (fig.1A) is to hybridize a labelled DNA fragment to the complete set of oligonucleotides immobilized on a plate as dots forming a 2-D matrix. The oligonucleotides corresponding to positively hybridizing dots of the matrix are considered to represent all the subsequences of the fragment. The reliability of information obtained during this step depends on simultaneous discrimination between perfect duplexes and defective ones (that are shorter duplexes, duplexes with mismatches and loops, etc.) in every dot of the oligonucleotide matrix. However, some types of mismatches are rather stable [3], their stability depending on the surrounding sequences. Therefore, to discriminate them simultaneously is not a trivial task.

We are looking for the discriminative hybridization conditions in model experiments using an exemplary set of immobilized oligonucleotides. Short oligonucleotides (8–15 bases long) are immobilized as dots on a glass plate. After the hybridization with a radioactively labelled probe, the plate is subjected to a series of washes at stepwise increased temperatures. The hybridization signals are measured for each dot after each washing step providing a set of kinetic dissociation curves that characterize duplex stabilities.

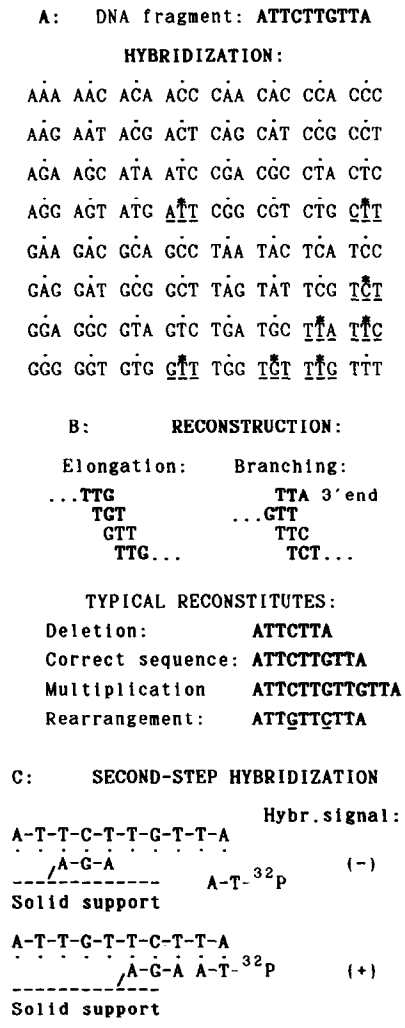

Fig.1. A schematic representation of the proposed sequencing procedure. A, B and C are discussed in sections 3.1, 3.2 and 3.3, respectively. A sequenced DNA fragment, octa- and heptanucleotides are represented for simplicity by deca-, tri- and dinucleotides, respectively. (C) The dinucleotide AT represents a very short oligonucleotide for second step hybridization.

Fig.2 shows the curves for immobilized oligonucleotides of 8, 9 and 10 bases long. The curves are well-separated from each other, thus showing the possibility to discriminate at least shorter duplexes. Unfortunately, in standard hybridization conditions this is not true for all duplexes.

It is worth mentioning that the measuring procedures described above allow an elegant technological realization with both radioactivity and fluorescence-labelled probes. The distribution of the probe on the plate can be converted into an optical image and transferred directly to a computer after each wash.

The use of dissociation curves provides the means to 'normalize' the entire matrix of oligonucleotides by the hybridization with the mixture of all the oligonucleotides of the complete set. As far as a perfect duplex is formed in each dot of the matrix, the stepwise washing provides a computer with the complete set of dissociation curves. With the matrix having been thus normalized, a positive hybridization signal at a certain dot of the oligonucleotide matrix is defined as the occurrence of hybridization events which fit a perfect duplex dissociation curve being saved in the computer memory for that particular dot. Such a procedure makes the method independent of the relative stabilities of different perfect duplexes. The use of dissociation curves also enables one to multiply minute differences in duplex stability at the expense of washing off a high proportion of hybridized probe.

The technology of synthesis and immobilization of multiple oligonucleotides could be greatly simplified if they were directly synthesized on the 2-D support. The procedure might be carried out by a printer-like device capable of sampling each of 4 nucleotides into given dots of the matrix. It is worth noting that methods of direct synthesis of immobilized oligonucleotides (without detachment–attachment step) are already available [4].

3.2. *Reconstruction*

A program for the reconstruction of a DNA fragment, providing a list of its subsequences, is based on the following algorithm (fig.1B). The reconstruction is started from any positively hybridizing octanucleotide. The next step reiterated many times is to find a successive

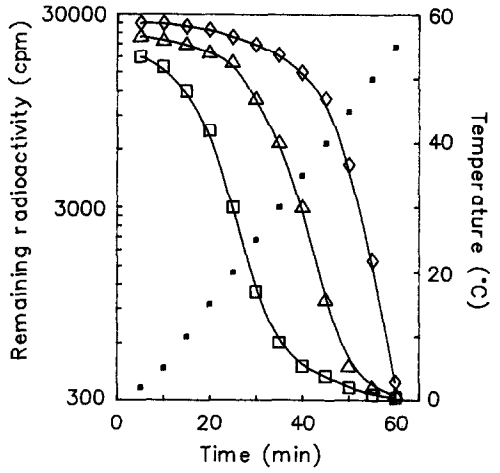

Fig.2. Dissociation curves of immobilized duplexes. A labelled heptadecanucleotide (^{32}P-GTAAAACGACGGCCAGT) was hybridized to oligonucleotides: CCGTCGTT (□), GCCGTCGTT (△), and GGCCGTCGTT (◇), immobilized on a single glass plate. Temperatures of washes are indicated as (■).

positively hybridizing octanucleotide with 5′ heptanucleotide identical to 3′ heptanucleotide of the preceding one, each iteration elongating the reconstructed sequence by one nucleotide in the 3′ direction. An analogous procedure enables one to move in the 5′ direction.

The reconstruction in 3′ direction is considered to be completed if there is no positively hybridizing octanucleotide, with its 5′ heptanucleotide being identical to the 3′ heptanucleotide of the reconstitute. The 5′ reconstruction is completed analogously. Thus the 5′ and 3′ ends of the fragment can usually be defined without any additional information.

A number of variants of elongation are possible if there are several positively hybridizing octanucleotides that differ only in the 5′ or 3′ nucleotide (this situation ensues if at least a heptanucleotide is repeated several times in a DNA fragment). At these points (so-called 'branching points') the alignment should follow each branch, thus generating several reconstitutes (in fig.1B dinucleotide TT represents a triple branching point). The differences between the correct variant and incorrect ones fall into 3 groups: deletions, multiplications and rearrangements. The deleted or multiplied sequences are embedded between repeated heptanucleotides (e.g. TTGTT in fig.1B).

One of the flanking heptanucleotides is also deleted or multiplied. Rearrangements occur for pairs of sequences embedded between identical pairs of heptanucleotides (e.g. TTCTT and TTGTT, fig.1B).

Those reconstitutes that contain deletions are readily recognized because they do not involve some positively hybridizing oligonucleotides. Multiplications may be identified either by the increased length of the corresponding fragment or by the fact that they contain extra copies of some sequences, provided the hybridization measures the number of oligonucleotide copies within the DNA fragment. The main difficulty is to identify the reconstitutes containing rearrangements. This can be achieved by different means, for example, by a second step of continuous stacking hybridization.

3.3. *Continuous stacking hybridization*

As the number and density of branching points increase with the length of DNA fragment, the approaches outlined in the previous paragraph become insufficient for unambiguous reconstruction. The length of a conseq can be further increased if the uncertainty at branching points is decreased by an additional continuous stacking hybridization of short oligonucleotides.

Continuous stacking hybridization of two duplexes (fig.3A) results in their mutual stabilization, possibly due to an extra stacking interaction, the stabilization being approximately equal to the addition of a base pair [5]. This effect is readily observed in our system for heptanucleotides (fig.3B) and we currently try to extend the procedure to penta- and tetranucleotides.

Fig.1C shows a scheme for the discrimination of a reconstitute with a rearrangement. Here, the rearrangement appears due to a triple repetition of dinucleotide TT. The second step hybridization of a labelled dinucleotide AT is performed in the AGA dot. Negative hybridization signal means, that the upper reconstitute is the correct one, as it is this reconstitute of the two, which does not support continuous stacking. If the reconstitute shown at the bottom was the correct one, the signal would be positive.

Actually, the dots of the octanucleotide matrix and the corresponding oligonucleotides for the second step hybridization should be chosen by a computer according to the results of first step

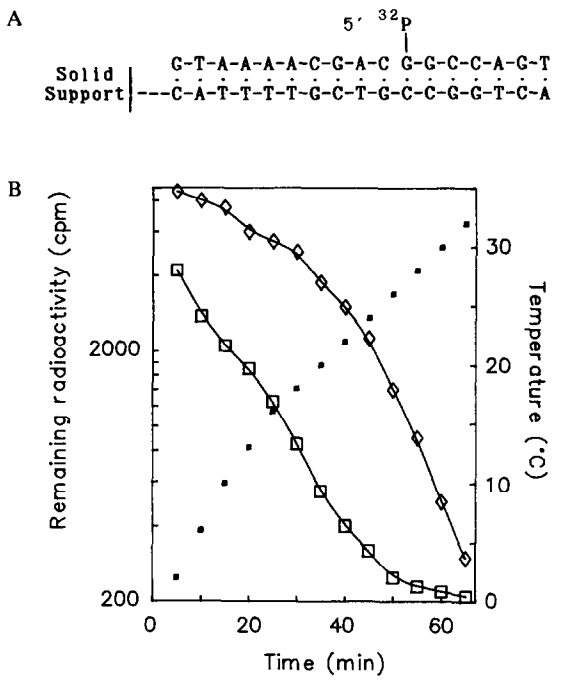

Fig.3. (A) A scheme of solid-supported hybridization of a heptanucleotide in continuous stacking with a decanucleotide. (B) Dissociation curves for the duplex shown in A (◇) and for the same duplex except the decanucleotide was omitted (□).

hybridization. This scheme implies that the complete set of non-immobilized short oligonucleotides is available.

3.4. *Conclusions*

A computer simulation of random sequences has shown that while one-step hybridization with the complete set of octanucleotides is sufficient to determine unambiguously the sequences of about 80% DNA fragments 200 bases long, a two-step hybridization enables one to sequence the same proportion of fragments of 350 bases long or greater than 90% of fragments of 200 bases long. It should be stressed that, even if a sequence cannot be reconstructed, a computer usually proposes a small number of variants which could be distinguished by different means.

Hence we consider the feasibility of the proposed method for determination of sequences approximately 400 bases long. The simultaneous reconstruction of overlapping fragments will further increase the length of sequences reconstructed

unambiguously. Although the method appears to be very laborious for sequencing a few DNA fragments, its simplicity for automation could lead to great advantages for large scale experiments necessary for the entire genome sequencing.

Acknowledgements: The authors are grateful to Dr S.K. Vasilenko for communicating his method of immobilization prior to publication and for helpful discussions.

REFERENCES

[1] Lysov, Yu.P., Florentiev, V.L., Khorlyn, A.A., Khrapko, K.R., Shick, V.V. and Mirzabekov, A.D. (1988) Dokl. Akad. Nauk SSSR 303, 1508–1511.
[2] Drmanac, R., Labat, I., Brukner, I. and Crkvenjakov, R. (1989) Genomics 2, 143–155.
[3] Ikuta, S., Takagi, K., Wallace, R.B. and Itakura, K. (1987) Nucleic Acids Res. 15, 797–811.
[4] Duncan, C.H. and Cavalier, S.L. (1988) Anal. Biochem. 169, 104–108.
[5] Springgate, M.W. and Poland, D. (1973) Biopolymers 12, 2241–2260.