Research Report

Computational Methods for Sequence Mapping of Large Combinatorial Libraries and Deduced Sequence Signatures

BioTechniques 34:1280-1286 (June 2003)

Valérie Abécassis, Lawrence Aggerbeck, Gilles Truan, and Denis Pompon

Centre National de la Recherche Scientifique, UPR 2167, Gif-sur-Yvette, France

ABSTRACT

Here we describe a computational approach for the high-throughput sequence mapping of combinatorial libraries obtained by DNA shuffling. Original algorithms and their software implementation were developed for the automated and reliable analysis of hybridization data of differentially labeled oligonucleotide probes with PCR products spotted on DNA microarrays. This novel approach allows a context-dependent sequence attribution tolerant to fluctuations in experimental conditions and is well adapted to hybridization signals of variable qualities resulting from high-throughput PCR amplification from colonies. In addition, the analysis permits the calculation of sequence signatures that are characteristic of combinatorial library structure, defects, and diversity. The approach is of interest for the characterization and the equalization (library reduction to nonredundant structures) of combinatorial libraries involved in directed evolution and could be extrapolated to high-throughput polymorphism analysis.

INTRODUCTION

Combinatorial libraries built by DNA shuffling of sequence variants or gene families (1,2) generally suffer from strong recombination biases (3,4) that dramatically reduce sequence diversity, particularly when the similarity between the recombined sequences is limited (5,6). For the selection of useful recombinants, a time-consuming and costly functional screening of large populations is required (7,8). Consequently, a high-throughput strategy enabling sequence mapping and detection of redundancy (9,10) would allow the equalization of combinatorial libraries (conversion of a large biased library to a smaller one containing only unrepeated sequences) and would critically improve the functional screening processes.

Hybridization of short radiolabeled oligonucleotides on target cDNA spotted on filters was previously reported (11–14). Such experiments led for each clone in the library to a characteristic fingerprint that is used for cluster analysis (14–16). This method was used for a large range of applications (10,17–20). We reported a similar method for the mapping of sequence distributions in combinatorial libraries of mosaic genes and leading to the determination of a sequence signature specific of the tested library (3). However, this microarray approach was not found to be easy to implement with large libraries. This task can be better performed using competitive hy-

bridization of differentially labeled oligonucleotide probes on glass-supported microarrays. An adapted data-processing method involving the transformation of fluorescence signal intensities in polar coordinates for sequence assignment was also described in an accompanying paper. However, if such simple processing was well adapted for good-quality data, it was insufficient for bad-quality signal encountered in some probe sets.

A more general procedure for processing data for sequence assignment and determination of the sequence signatures of libraries is described. The data processing involves (i) two filtering steps designed to eliminate non-pertinent data, (ii) the calculation of sequence discrimination factors derived from raw fluorescence data, and (iii) a modeling step that is used to adapt sequence discrimination criteria to the particular features of each slide. The algorithms presented allow robust and automated analyses of very large sets of data. Deduced sequence maps for each clone are then used for a second round of calculation to determine sequence signatures that describe the types and frequencies of mosaic structures found in the library.

MATERIALS AND METHODS

Library Constructions

The construction of libraries A–C is described in the accompanying paper (Abécassis et al., Microarray-Based

Method for Combinatorial Library Sequence Mapping and Characterization) in this issue of *BioTechniques*.

Experimental Data

PCR amplifications of the cDNA inserts of 384 randomly picked clones of each library were performed. Purified DNA products from the three libraries were spotted in quadruplicate onto coated glass slides and hybridized sequentially with the six different probe pairs. Spot fluorescence was determined on the red and the green channels with an Axon scanner. The scanner output file (GenePix® Pro 3.0 software; Axon Instruments, Union City, CA, USA) was combined with the spotter output file (GeneTac™ G3 software; Genomic Solutions, Ann Arbor, MI, USA) and the microplate maps to associate each of the 4608 ($384 \times 3 \times 4$) spots in each slide with the corresponding clones and libraries. All chip-related operations were performed on the Gif-Orsay DNA Mi-

croarray Platform at Gif-sur-Yvette. Data were processed as described in this paper using Microsoft® Excel® worksheets and Visual Basic® routines covering all steps and offering visual interfaces for parameter adjustments.

RESULTS AND DISCUSSION

Experimental Data Set

Slides were printed with the three combinatorial libraries spotted in quadruplicate and were probed by sequential hybridization with six pairs of probes. Each probe set is constituted of a pair of CyTM3- and CyTM5-labeled oligonucleotides that target the two possible parental sequence types on a given sequence segment. Green and red fluorescence intensities were used to monitor hybridizations of the two parental probes (see Materials and Methods). Hybridization signals for PCR products from libraries A and C had similar fluorescence intensity distributions. They were used to illustrate the analysis of good-quality signals. In contrast, signals for library B were 10- to 50-fold lower and were analyzed separately to illustrate the case of a low signal that can result from limited PCR amplification or from spotting with imperfectly purified materials.

was used to filter out “atypical signals” (spots exhibiting high intensities on both channels and limited probe discrimination) and for sequence type attribution of regular signals. Sequence attribution based on R2 indexes was straightforward for the majority of the population of spots exhibiting high signals. Nevertheless, the modeling of the R2 index distribution in the global population was needed for a more reliable sequence attribution of spots exhibiting ambiguous signals (see latter).

Signal Processing and Filtering

Absolute values of hybridization signals were found useless for sequence type attribution because of the high dispersion of the hybridization signal from spot to spot in the library. Simple calculation of the ratios of intensities of green and red signals (described as R0 in Table 1, equation 3) also does not allow a reliable sequence attribution, as a smooth distribution curve for the ratio was observed, particularly with the library having low signals (Figure 2, panels A1 and A2).

The situation can be partially improved by preprocessing the data by calculating a derived parameter: the R1 index as described in Table 1, equation 4. This calculation depends on an arbitrary threshold parameter (Th) and allows the transformation of the broad distribution observed in Figure 2, panel A1, into a trimodal one (Figure 2, panel B1). However, with the low-signal library, the same calculation transforms the distribution of Figure 2, panel A2, into the distribution of Figure 2, panel B2, which is still useless for sequence attribution. This results from the high proportion of low signal spots for library B compared to a much higher signal for libraries A and C. To solve this difficulty, a filter was designed to filter out populations exhibiting non-significant signals. This can be simply done by defining a Lower Signal Threshold (LST) parameter and by filtering out populations of spots exhibiting signals on both red and green channels lower than LST. Provided with a suitable trial-and-error choice of the Th parameter, the distribution of the R1 index for this low-signal population exhibited a Gauss-shaped zero-centered frequency curve without shoulders (Fig-

Figure 1. Flowchart for the transformation of scanner data to sequence signatures. F and B refer to spot and local background fluorescence signals, respectively.

Synopsis of Data Processing

Figure 1 illustrates the flowchart for data processing and was converted into Excel worksheets and Visual Basic routines. In short, corrected fluorescence values (Table 1, equations 1 and 2) were used to calculate different index values. The first index (R1) was used in combination with absolute signal values to filter out spots exhibiting low-intensity signals on both channels and poor probe discrimination (mainly corresponding to free vectors and poor PCR amplifications). The second index (R2), also in combination with the absolute signals,

Microarray Technologies

ure 2, panels C1 and C2). Suitable adjustment of the Th and LST parameters can be further confirmed by checking that control spots corresponding to

cDNA-free plasmids all belong to the filtered low-signal subpopulation.

Once Th and LST parameters have been adjusted, the frequency curve for

the R1 indexes of the remaining population exhibits a more bimodal distribution (Figure 2, panels D1 and D2) both for low- and high-signal libraries. However, the distribution is still smooth and difficult to use for statistical analysis. To improve discrimination further, a new index R2 was defined (Table 1, equation 5). The function $R2 = f(FC_{635}, FC_{532})$ offers the interesting property of being continuous and monotonic for FC_{635} or FC_{532} values close to 0, including negative values of FC_{635} or FC_{532} . Slightly negative values can result, for example, from noise or local artifacts. The R2 values are strictly included in the $[-3; +3]$ interval, whereas the R1 values are unlimited. The R2 index, like the R1 index, is null when the 532- and 635-nm channels exhibit equal intensities. When R2 indexes were plotted versus the mean intensities of red and green channels for the full population (libraries A, B, and C together) (Figure 3A), three subpopulations were visualized: one corresponding to low signals, which is the same as the one shown previously by R1 index analysis but is now distributed over the entire vertical axis, one intermediate population characterized by medium or high signal intensities on both channels and low absolute values for the R2 indexes, and a major bimodal population centered on R2 values of + or -1. The low-signal population can be filtered out as previously described. To filter out the intermediate population, a new threshold parameter, Mixed Signals Threshold (MST), was introduced. Spots for which signal intensities on both channels were above MST were considered to belong to this intermediate population. This population is also characterized by a rather low (<0.5 – 0.7) absolute value of the R2 index and appeared as yellowish brilliant spots on slides. Restriction analysis and sequencing reveal that these atypical spots constitute a non-homogenous population mainly corresponding either to a mixture of sequences, probably arising from incomplete plasmid segregation in the libraries, or from PCR-related mutations.

Figure 2. Data processing. The left panels refer to cumulated libraries A and B (768×4 spots), whereas the right panel refers to library C (384×4 spots). (A) Frequency of observation of R0 index values for spots in a typical slide. The y-axis gives the number of observations of log-ratio values belonging to classes centered on the values given on the x-axis. Forty classes of equal size were used for analysis. (B) Same as in panel A for the frequency of observation of R1 index values for spots with a Th value of 200. (C) Same as panel B for the population of spots that exhibit a red or green fluorescence lower than 300 units with a Th value adjusted at 200. (D) Same as panel B, but for the filtered population (exclusion of the low signal and the atypical population). (E) Same as panel D, except that R2 index was used in calculations instead of the R1 index.

Table 1. Equation Definitions

Calculation of the fluorescence data

$$FC_{532} = F_{532} - B_{532} \quad (\text{Eq. 1})$$

$$FC_{635} = F_{635} - B_{635} \quad (\text{Eq. 2})$$

F532 or F635 are mean values of the median of fluorescence intensities of pixels of the quadruplicate for 532- or 635-nm channels, respectively; B532 and B635 correspond to the mean values of the median of local background of fluorescence intensities around spots.

Calculation of the log-ratio indexes

$$R0 = \log_{10} \left(\frac{FC_{635}}{FC_{532}} \right) \quad (\text{Eq. 3})$$

$$R0 = \log_{10} \left(\frac{FC_{635} + Th}{FC_{532} + Th} \right) \quad (\text{Eq. 4})$$

Th is the defined threshold value.

Calculation of the polar coordinate-based indexes

$$\left. \begin{aligned} R2 &= 3 - \frac{4}{\pi} \times \text{Atn} \left(-\frac{F_{532}}{F_{635}} \right) \text{ if } F_{532} \geq 0 \text{ and } F_{635} < 0 \\ R2 &= 3 + \frac{4}{\pi} \times \text{Atn} \left(-\frac{F_{532}}{F_{635}} \right) \text{ if } F_{532} < 0 \text{ and } F_{635} \geq 0 \\ R2 &= \frac{4}{\pi} \times \text{Atn} \left(\frac{F_{532}}{F_{635}} \right) - 1 \text{ in all other cases} \end{aligned} \right\} \quad (\text{Eq. 5})$$

Atn is trigonometric arc-tangent value (in radian unit) using the function definition suitable for the angular range $-\pi/2$ to $+\pi/2$.

Calculation of the sequence index (I_m)

$$I_m = \sum_{u=1}^N S_m(u) * 2^{(u-1)} \quad (\text{Eq. 6})$$

N is the number of probed sequence segments; u is the index corresponding to the probed sequence segment (1–6 here); $S_m(u)$ is the value 0 (parental type A) or 1 (parental type B) for sequence segment corresponding to the index u for the mosaic sequence m .

Calculation of the frequencies for mosaic classes $F(I_m)$

$$F(I_m) = \sum_{i=1}^{NP} \prod_{u=0}^{N-1} [S_m(u) * PA_m(u) + (1 - S_m(u)) * (1 - PA_m(u))] \quad (\text{Eq. 7})$$

N is the number of tested sequence segments; NP is the number of tested clones in the library; $PA_m(u)$ is the probability to have a parental type “A” at probed sequence u in the mosaic m .

Following the filtering out of the low-signal and atypical populations, the remaining “normal” population was characterized by a predominantly bimodal distribution of R2 index values (Figure 2, panels E1 and E2), each mode corresponding to one of the two

possible parental sequence types (*CYP1A1* or *CYP1A2*). The frequency curve of observed R2 indexes for this population can be easily fitted using a Gaussian curve decomposition (Figure 3, B and C). Sequence attribution did not usually require such decomposition

for most of the probe sets, as non-overlapping distribution curves with clear-cut modes were observed (Figure 3B). However, for low-signal libraries or for some particular probe sets (depending on the sequences), the distribution curve remained multimodal (Figure 3C), making the multiple Gaussian curve decomposition useful for sequence attribution in a probabilistic way based on modeling. All these data analysis processes have been included

in specific software allowing, in addition, easy user-interactive adjustment of all the parameters (Th, LST, and MST) through a graphical interface. Such software fully automates the conversion of raw data from the scanner to the final sequence map attributions.

Sequence Signature Analysis

The sequence signature analysis consists of the evaluation of the fre-

quencies of observation of all possible mosaic structures that can be identified in the combinatorial library (3). Such signatures are useful tools for robotic equalization of libraries (elimination of redundancies) and for structure-function relationship analysis at a population-wide level for the purposes of directed evolution. The establishment of detailed sequence signatures needs successful probing of a large number of sequence segments. Absence of sequence information even for only one segment per clone is enough to invalidate the calculations. Therefore, the “probabilistic” way of attributing sequence types by multiple Gaussian curve modeling is a significant advantage in that it allows the attribution of sequence probabilities to all segments and thus the calculation of signatures in all cases. The frequencies of observation $F(I_m)$ (defined in Table 1, equation 7) in the population of mosaic sequences characterized by sequence indexes of values, I_m , were calculated. The I_m sequence index is a descriptor (integer number) that encodes the map of a given mosaic as defined in Table 1, equation 6. The plot of $F(I_m)$ as a function of I_m and the corresponding cumulated frequency curve constitute the sequence signature of a library.

To evaluate how the sequence signatures of libraries could be affected by the quality of the hybridization signals during the chip analysis step, sequence attributions and signature calculations were performed, separately, on combinatorial libraries A, B, and C (Figure 4). While signal quality highly differed between libraries A and B, a similar signature was expected because of the very similar experimental protocol used to construct the two libraries. In contrast, libraries A and C were constructed in a different way, and, thus, different signatures were expected. The results presented in Figure 4 show that very similar signatures were effectively obtained for libraries A and B (Figure 4, A and B), which clearly differs from the one obtained for library C (Figure 4C). This demonstrates that the signature calculation is only marginally affected by signal quality over a large range (10- to 50-fold) of fluorescence intensities. In fact, the sequence signature can be decomposed into two major

Figure 3. Population discrimination and statistical modeling. (A) Population sorting based on discrimination indexes and signal intensities. The values for the R2 discrimination index are plotted versus the signal intensity modules (square root of the sum of the square of fluorescence intensities on each channel). Shaded areas indicate the limits of subpopulations according to the data analysis procedure described. (B and C) Multi-Gaussian decompositions of the frequency curves for the observed classes of values of R2 indexes for spot populations giving high signals (libraries A and B) and low signals (library C), respectively. Dots are experimental points.

Figure 4. Sequence signatures for combinatorial libraries deduced from chip data. Panels A–C correspond to sequence signatures obtained for libraries A, B, and C, respectively. The x-axis corresponds to the mosaic structure codes 0–63 (six sequence segments each belonging to two possible parental types were tested). Codes 0 and 63 correspond to parental sequences of P450 1A1 and 1A2, respectively. The y-axis represents the calculated frequency of observation for each mosaic structure.

components: the distribution of the frequencies of observation of each class of mosaic and the cross-correlations between the probabilities of observation of sequence segment associations in the combinatorial library. The first component defines the general signature shape and is characteristic of the statistical distribution of the parental sequence types along the sequence. Provided that assignment errors are randomly distributed, this distribution is statistically conserved, and the shape of the signature is, thus, maintained. In contrast, the cross-correlation coefficients for the probabilities of parental-type conservation between segments are extremely sensitive to assignment errors on individual mosaics. This results in the randomization of this second component and thus smoothing of the composite curve, without changing the shape, when bad-quality data are processed. This makes the signature extremely insensitive to fluctuations in experimental

condition and makes it a very stable description of the library structure.

CONCLUSION

A high-throughput tool for sequence mapping of combinatorial libraries is described. The experimental approach differs from the previously described DNA macroarray technique (3) by the use of competitive hybridization of differentially labeled short oligonucleotides on PCR products printed at high density on a glass slide. The calculation of a library sequence signature typically required analyses of large numbers of individual hybridizations to be of practical interest in the mapping of combinatorial libraries used in directed evolution techniques. Indeed, large numbers of clones must be characterized with a large number of sequence segments. The automated, although user-super-

vised, procedure described here provides a solution to several difficulties associated with such a high-throughput analysis (21,22), such as the heterogeneity of the hybridization signal that is associated with variability in the PCR amplification processes performed on colonies, the presence of unexpected mutations leading to non-standard hybridization signals, and the limited discrimination of some probes. In addition, the probabilistic approach used for sequence attribution takes into account slide-to-slide signal fluctuations by analyzing the behavior of individual spots in light of the behavior of the entire spot population on the same slide. This is achieved by data filtering to eliminate non-pertinent data and by the use of both modified log-ratio and polar representations to calculate discrimination indexes suitable for Gaussian distribution curve modeling. As judged by sequence signature conservation in two experiments in which the experimental signal differed by more than one order of magnitude, the resulting sequence maps were found to be extremely robust to changes in experimental conditions. Sequence signatures are potentially interesting for reasons other than those illustrated for the characterization of the diversity and the defects of combinatorial libraries. Particularly, the variation of sequence signatures of combinatorial libraries in response to functional selection pressures is a potential innovative approach for the identification of critical sequence segment associations for directed evolution purposes (3,4). Sequence signatures are also a basic tool necessary for performing and evaluating robotic equalization (the reduction of the population to nonredundant structures) or volunteer biasing (selection of a particular subset of structures) of combinatorial libraries.

ACKNOWLEDGMENTS

We wish to thank Philippe Urban for the careful proofreading of the manuscript. The work was supported by a grant from CNRS for DNA microarrays, from the GIP Hoechst Marion Roussel for combinatorial libraries, and for core equipment by the Gif-Orsay

DNA microarray platform. V.A. is a pre-doctoral fellow supported by a fellowship from the Ligue Nationale Contre le Cancer.

REFERENCES

1. Stemmer, W.P. 1994. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA* 91:10747-10751.
2. Cramer, A., S.A. Raillard, E. Bermudez, and W.P. Stemmer. 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* 391:288-291.
3. Abécassis, V., D. Pompon, and G. Truan. 2000. High efficiency family shuffling based on multi-step PCR and in vivo DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2. *Nucleic Acids Res.* 28:E88.
4. Joern, J.M., P. Meinhold, and F.H. Arnold. 2002. Analysis of shuffled gene libraries. *J. Mol. Biol.* 316:643-656.
5. Gibbs, M.D., K.M. Nevalainen, and P.L. Bergquist. 2001. Degenerate oligonucleotide gene shuffling (DOGS): a method for enhancing the frequency of recombination with family shuffling. *Gene* 271:13-20.
6. Kikuchi, M., K. Ohnishi, and S. Harayama. 2000. An effective family shuffling method using single-stranded DNA. *Gene* 243:133-137.
7. Arnold, F.H., P.L. Wintrobe, K. Miyazaki, and A. Gershenson. 2001. How enzymes adapt: lessons from directed evolution. *Trends Biochem. Sci.* 26:100-106.
8. Cohen, N., S. Abramov, Y. Dror, and A. Freeman. 2001. *In vitro* enzyme evolution: the screening challenge of isolating the one in a million. *Trends Biotechnol.* 19:507-510.
9. Eickhoff, H., J. Schuchhardt, I. Ivanov, S. Meier-Ewert, J. O'Brien, A. Malik, N. Tandon, E.W. Wolski, et al. 2000. Tissue gene expression analysis using arrayed normalized cDNA libraries. *Genome Res.* 10:1230-1240.
10. Radelof, U., S. Hennig, P. Seranski, M. Steinfath, J. Ramser, R. Reinhardt, A. Poustka, F. Francis, and H. Lehrach. 1998. Preselection of shotgun clones by oligonucleotide fingerprinting: an efficient and high throughput strategy to reduce redundancy in large-scale sequencing projects. *Nucleic Acids Res.* 26:5358-5364.
11. Strezoska, Z., T. Paunesku, D. Radosavljevic, I. Labat, R. Drmanac, and R. Crkvenjakov. 1991. DNA sequencing by hybridization: 100 bases read by a non-gel-based method. *Proc. Natl. Acad. Sci. USA* 88:10089-10093.
12. Meier-Ewert, S., E. Maier, A. Ahmadi, J. Curtis, and H. Lehrach. 1993. An automated approach to generating expressed sequence catalogues. *Nature* 361:375-376.
13. Drmanac, R., S. Drmanac, Z. Strezoska, T. Paunesku, I. Labat, M. Zeremski, J. Snoddy, W.K. Funkhouser, et al. 1993. DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing. *Science* 260:1649-1652.
14. Milosavljevic, A., Z. Strezoska, M. Zeremski, D. Grujic, T. Paunesku, and R. Crkvenjakov. 1995. Clone clustering by hybridization. *Genomics* 27:83-89.
15. Herwig, R., A.J. Poustka, C. Muller, C. Bull, H. Lehrach, and J. O'Brien. 1999. Large-scale clustering of cDNA-fingerprinting data. *Genome Res.* 9:1093-1105.
16. Hartuv, E., A.O. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, and R. Shamir. 2000. An algorithm for clustering cDNA fingerprints. *Genomics* 66:249-256.
17. Craig, A.G., D. Nizetic, J.D. Hoheisel, G. Zehetner, and H. Lehrach. 1990. Ordering of cosmid clones covering the herpes simplex virus type 1 (HSV-1) genome: a test case for fingerprinting by hybridization. *Nucleic Acids Res.* 18:2653-2660.
18. Drmanac, S. and R. Drmanac. 1994. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *BioTechniques* 17:328-336.
19. Meier-Ewert, S., J. Lange, H. Gerst, R. Herwig, A. Schmitt, J. Freund, T. Elge, R. Mott, et al. 1998. Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Res.* 26:2216-2223.
20. Drmanac, S., D. Kita, I. Labat, B. Hauser, C. Schmidt, J.D. Burczak, and R. Drmanac. 1998. Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nat. Biotechnol.* 16:54-58.
21. Eisen, M.B. and P.O. Brown. 1999. DNA arrays for analysis of gene expression. *Methods Enzymol.* 303:179-205.
22. Schuchhardt, J., D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, and H. Herzel. 2000. Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* 28:E47.

Received 3 December 2002; accepted 20 February 2003.

Address correspondence to:

Dr. Denis Pompon
Centre de Génétique Moléculaire
Centre National de la Recherche Scientifique
UPR 2167
91190 Gif-sur-Yvette, France
e-mail: pompon@cgm.cnrs-gif.fr