DOE/ER/62923

DE-FG02-00ER62923
Final Report: Time-Resolved Sequence Analysis on High Density Fiberoptic DNA Probe Arrays
Tufts University
Principal Investigator: David R. Walt
Title: Time-Resolved Sequence Analysis on High Density Fiberoptic DNA Probe
Reporting period: March 1, 2000-February 28, 2002
Authors: David R. Walt and Kyong-Hoon Lee

Abstract

A universal array format has been developed in which all possible n-mers of a particular oligonucleotide sequence can be represented. The ability to determine the sequence of the probes at every position in the array should enable unbiased gene expression as well as arrays for *de novo* sequencing.

Results

Creating a DNA fiber optic microsphere array

A 500-μm diameter imaging fiber containing 10,000 individually clad optical fibers was used as the substrate. Dipping the tip of the imaging fiber into a buffered acid solution simultaneously creates 10,000 wells due to the difference in etching rates of the core and cladding material. The diameter of each well created is 3.5 μm and the depth is 3 μm, controlled by the amount of time in the acid solution. The microsphere array is created by placing a 50 nL drop of water containing 3.2 μm diameter beads onto the etched tip of the imaging fiber. As the water evaporates, the beads settle individually into the wells and excess beads are removed by gently wiping the fiber tip. The beads residing in the wells are independently and simultaneously addressed using an imaging system with a CCD detector.

The DNA microsphere array was prepared as described previously. Briefly,

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

single strand 3'-amino terminated oligonucleotide probes are attached to the microsphere via a polyethyleneimine linker. The functionalized beads are placed in the wells of the fiber optic substrate. Upon hybridization and excitation, the bound fluorescently labeled targets emit a fluorescent signal. Regeneration using 90% formamide removes the hybridized target without compromising the array's integrity.

Demonstrating the fiber optic microsphere sequencing array

To demonstrate the concept of the fiber optic sequencing array, we synthesized 15 probe sequences with an 8-mer registration sequence and a 4-mer ID segment. The 4-base ID segment model sequence is AAAA. The probes are named according to the ID segment base position and variation from AAAA.

The registration sequence consists of an 8 base sequence rich in GC content. The length and content were selected for hybrid stability. Two registration sequences were used, both generating completely specific identification. The first registration sequence, DC1, was arbitrarily designed maintaining a high GC content: GCG GTC CC. We were able to characterize the ID segment effectively and specifically, however, we noticed a decrease in complementary hybridization in the stringent hybridization conditions. The second registration sequence, abbreviated WT, was selected from a wild type K-ras oligonucleotide used in previous work and known for its stability in stringent hybridization conditions: GGA GCT GG. A 20% formamide buffer solution provided the stringency for consistent single base mismatch discrimination while maintaining complementary hybridization. Using this registration sequence, hybridization of a perfect complement was stable in the formamide buffer.

Combinatorial synthesis of decoding targets

The randomly distributed probes are mapped using pooled decoding solutions. Each decoding solution contains 4^n synthesized targets complementary to every possible probe in the array. The ID segment used in this demonstration was a 4 mer therefore; the decoding target solution contained 4^4=256 targets. Each target in the decoding solution (decoding target) includes the complement to the registration sequence and an ID segment. Decoding solutions are prepared using a mix and split combinatorial synthesis scheme. Each base is assigned a fluorescent tag and each decoding target is labeled according to the particular base identity that has been specified. Four separate syntheses are conducted for each position of the ID length. The four targets for a defined position are then mixed providing one decoding solution for each position of the ID segment. The identity of the probe sequences and a map of the array are then determined using only *n* decoding solutions.

For all decoding targets, the registration sequence is the same (example CGC CAG GG) and a defined base is synthesized at a defined position of the ID segment (example T at position 9 CGC CAG GG**T**, this synthesized target is referred to hence as 9T). The remaining positions of the decoding target are programmed to receive equal amounts of each of the four bases, resulting in each of the four bases appearing at every position in the sequence other than the defined base. For example, after the T is placed at position 9, the remaining three positions 10-12 are combinatorially randomized with all four bases such that the resulting pool contains T at position 9 and X at positions 10-12 where X=A, C, G, T (CGC CAG GGT XXX). Once the *n* length of the ID segment is achieved (in the present example the ID segment is 4 bases in length), the decoding target

is terminally labeled according to the defined ID base (for this example T is labeled with Cy5). According to the example, the decoding target synthesized (9T) includes equal concentrations of 64 sequences with each sequence terminally labeled with Cy5. The remaining position 9 decoding targets are then separately synthesized with position 9 defined with C (9C), G (9G), or A (9A), with the other positions randomized, and each labeled with a specified dye. The four targets (actually 4X64=256), 9G, 9C, 9T and 9A are then combined creating a decoding solution for position 9 containing 4^4 or 256 different targets. The decoding solution for each position is used to determine the base in the ID segment on each microsphere in the array. It is important to note that for a given registration sequence and a given probe length, the decoding solution for a particular position is universal.

Due to synthesis constraints in building from the 3' end and labeling the 5' end of the target, we selected one base (T) to initiate the synthesis of all targets. Therefore, our decoding solutions were reduced in complexity by 4 and contained 256/4=64 different sequences. The 9T target consists of the following sequence: 5'-CCT CGA CCT XXT-3' yielding 4^2 different bases. When all position 9 targets are combined (9T, 9C, 9C, 9G) we have 64 sequences. The required final concentration of our target for providing indisputable signals from the complementary base in less than 5 minutes was 10 μM.

Position Specificity

To determine the integrity of the discrimination at each position of the ID segment, the DC1 probe microsphere GCG GTC CCA AAA was placed alone in a fiber optic well array. Four targets were synthesized, the perfect complement and three

containing a single base mismatch (e.g. a C in place of a T) in each of the four positions of the ID segment. The complementary target was labeled with Cy5 and the three targets containing a single base mismatch were labeled with fluorescein. Competitions between the complement and each of the four single base mismatch containing targets were conducted. Using the complementary target and one of the single base mismatches in the hybridization solution, we noted a decrease in specificity as the competitive target mismatch moved from position 9 to position 12 (position 12 is farthest from the bead and registration sequence). Although there was a decrease in specificity, the signal from the complement remained the dominant signal. Using the more complex decoding solution (containing 256 sequences and four dye labels), we are also able to elicit a correct identification of the base at position 12 by noting the dominant signal.

Decoding

The decoding solution provides a competition among the 256 possible sequences of the ID segment. All targets are labeled and can potentially produce a hybridization signal. This complex solution creates higher background levels from all non-complementary targets. Conducting the competitive assay in buffer and in buffer containing 20% formamide, we were able to identify the base at positions 9 ans 10 of the ID segment on each bead by the dominant fluorescence signal resulting from hybridization of the fully complementary target sequence. With the 20% formamide buffer, we are able to reduce all background signals to a level substantially below the signal generated from the perfect complement's signal. Position 12, the terminal position, was difficult to decode in buffer alone. Complete and indisputable determination of the

base identity at the terminal position, however, was always achieved using the 20% formamide buffer.

Dyé labels

Cyanine dyes have much greater quantum efficiency than fluorescein. The acquisition times required to generate similar signals from the four labels were adjusted to account for this difference. The acquisition time used at the fluorescein wavelength was 10 times greater than the acquisition times at the cyanine wavelengths.

We also noted that the fluorescein phosphoramidite label generated significantly lower signals than the fluorescein-streptavidin label. This low fluorescence signal was not related to hybridization since the same result was seen when using fluorescein to specify different bases. The targets for the WT registration were synthesized with fluorescein specified for the known base T and the targets for the DC registration were synthesized with fluorescein specified for the known base C. The results show that the effect from fluorescein phosphoramidite is not base sensitive. To resolve this situation, we increased the concentration of all fluorescein labeled targets three times in the decoding solutions and were clearly able to identify all four bases of any ID segment.

Increasing length of ID segment

We were also interested in determining if this decoding strategy could be applied to more complex target solutions. We selected an ID length of 12 (producing a 20 mer sequence with the 8 mer registration sequence). We selected two probes to interrogate positions 13 and 19 of the 20 mer. The two probe sequences were 5'-**GG AGC TGG**

AAA AGA AAA AAA-3' named 13G and 5'-**GG AGC TGG** AAA AAA AAA ACA-3' named 19C (with the registration sequence and difference in bold type).

The decoding solutions were combinatorially synthesized as described above. The final decoding solution contained 4^{12} ($1.68x10^7$) different targets with only one of the targets being the perfect complement to a probe. An intensified CCD camera and an increase in the target concentrations to 125 □M were required in order to obtain a hybridization signal. We were able to see the hybridization signal after 20 minutes of incubation.

Hybridization conducted in the 20% formamide buffer solution provided the specific identity of the base at position 13 of each probe determined by the highest signal from the four labels. We were unable to accurately identify the base at position 19 due to the low fluorescence values and decreased specificity. This failure may stem from both the complexity of the decoding solution as well as the lower fidelity of hybridization to the more terminal bases of the probe sequence.

Longer Sequesnces-Additionally, we investigated whether this decoding strategy could be applied to a longer sequence (a 20-mer sequence = 8-mer registration sequence + an ID length 12-mer) with more complex target solutions (4^{12} =$1.68x10^7$ targets, the total concentration = 125 μM). Two probes were used to interrogate the 13^{th} and the 19^{th} positions of the 20-mer, of which the sequences were 5'-**GG AGC TGG** AAA AGA AAA AAA-3' named 13G and 5'-**GG AGC TGG** AAA AAA AAA ACA-3' named 19C (with the registration sequence and single base difference in bold type). After 20 minutes of incubation, we were unable to accurately identify the base at the 19^{th} position due to the low fluorescence values and decreased specificity. Difficulties decoding this position

stemmed from both the complexity of the decoding solution as well as the lower hybridization fidelity associated with terminal bases.

To address this problem, we attempted to enhance the signal strength without modification of our system's general experimental conditions. Previous experiments have demonstrated the number of microbeads loaded on the distal tip of a fiber directly affects the signal. For example, a complete hybridization of 1,000 total target molecules (approximately 10 μL of a 1 fM solution) to an array containing 1,000 beads would provide a single target molecule per bead. If the same experiment was performed with only ten beads in the array, the average number of target molecules per bead would now be 100, and a substantially greater signal would be obtainable. Using the 12-mer probe with a 4-mer ID segment of 9C10G11A12A, this concept was applied to two experiments. In the first experiment, an array comprised of ~2,000 beads was fabricated and hybridized for 5 minutes with the target decoding solution containing 20 % formamide. For the internal base positions (9C and 10G), all randomly selected beads gave correct decoding responses. The terminal positions (11A and 12A), however, could not correctly identify the polymorphic base. A second experiment was performed with a reduced number of beads (*ca.* 70). In this case, every randomly selected bead enabled correct identification of all the four polymorphic probe sequences, regardless of their positions. Previous experiments have also demonstrated that extension of the hybridization time leads to greater signal. Longer hybridization times enable more complementary, fluorescently-labeled target molecules to bind to the microsphere probes.

A series of decoding experiments with extended hybridization times and minimal bead numbers was performed using the 20-mer probe with mutation at 19C. Combining

and 19AG) position were prepared. The three target decoding solutions for interrogating each polymorphic position were prepared in the 20 % formamide buffer.

As a preliminary step prior to multiplexed assays, a decoding experiment for each probe was separately performed to find out the optimal hybridization time. To identify the 9th position (9WT, 9TA, 9TG, and 9TC), 5 minute hybridization times were enough to correctly identify the polymorphic positions. As the polymorphic position became more terminal, the experimental hybridization times were increased, requiring 10 minutes for interrogating the 15th position, and 20 minutes for interrogating the 19th position.

Combinatorial 12-mer ID sequence synthesis

We developed a pilot run of combinatorial DNA synthesis to fabricate every possible 12-mer ID sequence. The combinatorial sequences were made in parallel by a modified split-and-mix protocol using standard solid-phase phosphoramidite chemistry methods. The synthetic product will be coupled to microspheres, post-fabrication, such that each microsphere has thousands of replicates of one unique sequence on its surface. The DNA-modified beads would then be used in a gene expression profile for specific, unbiased gene decoding.

Discussion

Previously, we conducted single base mismatch experiments with a wild type (WT) K-ras oligonucleotide. The hybridization signal of the perfect complement remained constant at room temperature and at 52°C while the signal due to hybridization of target with a SNP dropped to baseline at 52°C. Other stringency conditions such as salt

or formamide in the buffer solutions affect hybridization melting curves. We found that adding 20% formamide to the buffer provided completely specific point mutation discrimination without compromising the integrity of the complementary hybridization. Both heat and the formamide buffer provided similar results and we elected to use the formamide buffer for ease in assay technique.

For hybridization efficiency, a GC rich 8 mer registration sequence was employed. We first randomly selected a sequence presuming that this sequence would provide a strong binding force for the registration segment of the targets while not compromising SNP discrimination. Initial experiments confirmed the specificity of the targets using this sequence, however, a significant decrease in complementary hybridization in the 20% formamide buffer solution was noted. In previous work, we used a 15-mer K-ras sequence to demonstrate the single base mismatch selectivity using the fiber optic microsphere array. This experiment produced no change in the complementary hybridization signal in the SNP discrimination buffer. We selected an 8 mer segment of this K-ras sequence for use as the registration sequence and found more reproducible signals from targets hybridized in buffer and formamide buffer solution.

The difference between the complementary base's hybridization signal and the mismatch hybridization signals depends on the labeling dye and sequence position. Adjusting acquisition times and labeled target concentrations enabled us to generate signals in the same range for all dyes. We were able to create a protocol that provided correct identification of any base at any given position. The ideal situation for the ID of the array is to have targets with labels of similar quantum efficiency and requiring no secondary reaction. This will require replacing the fluorescein label and the biotin label.

Other available phosphoramidite dyes spectrally overlap Cy5 and Cy3. Synthesizing the decoding target with an amine end-label would enable the use of amine reactive dyes. There are many amine reactive dyes that can be used in conjunction with Cy5 and Cy3 without interference.

Decoding the longer, 12-mer ID sequences required an intensified CCD camera and a much higher concentration of target solution due to the solution's complexity. We were able to accurately determine the identity of the base in the center of the probe, position 13, under these conditions. Considering the complexity of the solution, 1.68×10^7 targets in a concentration of 125 □M with a volume of 30 □L, the final concentration of the complementary target was 7.4 fM.

DNA hybridizations are dependent upon the sequence (GC content, sequence length, and secondary structure), as well as the target concentration. Our approach to successfully decode an increased ID segment length involved increasing the local target concentration by minimizing the number of beads in the array, and increasing the hybridization times.

Decoding terminal probe positions by reducing the number of beads in the array provides enhanced local concentrations of target molecules. The successful decoding of the 4-mer ID sequence with fewer bead numbers supports this concept. Further experiments involving the 12-mer ID sequence, maintained total bead numbers below 10. Fewer beads were necessary because the subsequent total available target numbers complementary to the longer ID sequence (7.4 fM) is less than those for the 4-mer ID sequence (9.8 nM) by several orders of magnitude.

For the multiplexed assays, the sequences of the 12-mer ID segments were based upon the wild type K-ras gene. Hybridization times were reevaluated, based on the total 20-mer sequences, because the longer ID segments would have a greater influence on the overall hybridization kinetics. The longer ID sequence compromises the importance of the registration sequence design, whereby the 4-mer ID sequence had hybridization kinetics influenced more by the thermodynamics of the registration sequence. Considering that a prolonged hybridization time may increase non-specific adsorption of highly concentrated DNA targets, the design of an energetically favorable registration sequence can affect reproducibility.

We have investigated implementation of a reliable decoding system for a microsphere-based oligonucleotide array. In summary, the terminal positions of a given probe with a smaller complementary target concentration could be better identified by minimizing the number of beads and by extending the hybridization times. The overall results emphasize the importance of the proper control of the concentration of target molecules delivered to the corresponding probe, and demonstrate decoding of longer ID sequences is possible. Future work will involve reproducible decoding of the 19^{th} and 20^{th} positions, and applying this method to an array containing combinatorial ID sequences. Synthesizing targets with universal bases in the randomized positions would greatly simply the decoding target solutions. The specificity of these targets and their ability to elicit sequence identification will be investigated in the future.

both approaches led to correct base identification at this position. Careful washing of the fiber tip after hybridization was essential to minimize non-specific adhesion of target DNA molecules.

Simultaneous identification of 4 sequences

To demonstrate the ability to multiplex this system, we first manually bundled four etched imaging fibers together, color coded them, placed one type of bead in each fiber substrate and placed the bundle on the imaging system. The tips of the four fibers bearing the microspheres were dipped into the decoding solutions (in the formamide buffer) and monitored simultaneously. Each of the four decoding solutions was assayed. The sequences on all four fibers were correctly identified.

Next, the 20-mer probes (8-mer registration + 12-mer ID sequences) with single base mutations were prepared and assayed simultaneously. The registration sequence was not changed (5'-**GG AGC TGG**-3') and ID sequences of a 12-mer length were variations of the K-ras wild type (WT) sequence used for our previous work. The sequence was modified in three positions such that each overall 20-mer probe has a different point mutation either at the 9^{th}, 15^{th}, or 19^{th} position. For example, the four probes used for interrogating the 9^{th} position were the K-ras WT gene (5'-**GG AGC TGG** T GGC GTA GGT AA- 3') and three others prepared by varying the 9^{th} position (underlined T) to A, G, or C. The probes were named 9WT, 9TA, 9TG, and 9TC, respectively, to denote the polymorphic base. In this manner, additional probes for multiplexing the 15^{th} (15WT, 15AT, 15AC, and 15AG) or the 19^{th} (19WT, 19AT, 19AC,