CHAPTER TWO EXCERPT FROM:

The Application of Functional Genomics, Systems Biology and Drug Development to the Study of Infectious Disease

by

Jingchun Zhu

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMEDICAL INFORMATICS

University of California San Francisco

Chapter 2 . ArrayOligoSelector (AOS)

ABSTRACT

The complete genome sequence of an increasing number of organisms is becoming available. To exploit these new resources for the purpose of developing whole genome microarrays, we developed a program, ArrayOligoSelector (AOS), to systematically design gene-specific long oligonucleotide probes for entire genomes. For each open reading frame, the program optimizes the oligonucleotide selection based upon several parameters, including uniqueness in the genome, sequence complexity, probe secondary structure, GC content, and proximity to the 3' end of the gene.

Using AOS, we designed a long oligonucleotide microarray for the complete genome of *Plasmodium falciparum*, the most deadly causative agent of human malaria. This malaria chip has been used to study the transcriptome of the parasitic intraerythrocytic developmental cycle and sequence variation between different *P. falciparum* strains.

AOS is an open source program and is freely available for public use at http://arrayoligosel.sourceforge.net. AOS has also been used by scientists all over the world to design whole genome microarrays for many other organisms such as *S. cerevisiae*, *M. musculus* and *H. sapiens*.

The first section of this chapter presents the AOS design algorithm. The second section is the documentation for the program.

Part I. AOS Algorithms

Background

Two important technological advances have been instrumental in transforming biological research from the study of a handful of genes at a time to the age of genomics. The first is whole genome shotgun sequencing and assembly that allows complete genome sequences be obtained much cheaper and faster. As a result, the number of fully sequenced genomes, strains or individuals has increased dramatically. The second advancement is DNA microarray, a powerful technology that allows simultaneous measurements of gene expression for every gene in a whole genome, which has been used to gain important insights into processes such as development, responses to environmental perturbations, gene mutation, and host response to pathogens, and cancer [1-5].

To efficiently transfer genome sequence resources to functional genomics using microarrays, a new kind of reagent - whole genome microarrays – is needed. The traditional method for constructing a whole genome array was to generate PCR products for every gene in the genome, a laborious and time-consuming process with various rates of success. This became extremely challenging for genomes with very high AT content such as that of *P. falciparum* (80% AT). In addition, PCR probes have difficulty distinguishing genes with a high degree of sequence similarity. Oligonucleotide probe based platforms provide an alternative that overcomes these disadvantages [6,7]. The use of synthetic oligonucleotide probes eliminates the need for PCR. By carefully selecting probes from the unique regions, this platform provides a means to readily distinguish between genes that have a high degree of sequence similarity and avoid other problematic regions such as the various types of repetitive sequences or secondary structures.

Several competing platforms for producing oligonucleotide-based microarrays have emerged, differing in probe length, number of probes required per gene, nature of the production processes, design customization, and cost [7]. Affymetrix (Santa Clara, CA) pioneered the commercial market by producing high density GeneChips using photolithography and solid-phase DNA synthesis, on which each gene is represented by a set (~20) of short oligonucleotides (20 –25mer) [8]. Alternative to using chromium masks in conventional photolithography, NimbleGen's (Madison, WI) maskless arrays (24 – 70mers) are produced by light-directed synthesis of oligonucleotides controlled by a digital micromirror device [9]. Other commercial platforms include Agilent's (Palo Alto, CA) microarrays produced by an inkjet printing technology that synthesizes 60mer probes [10], CodeLink Bioarray^{TM} (Amersham Biosciences, Piscataway, NJ) that uses a 3D polyacrylamide gel matrix as the slide surface for depositing 30mer oligonucleotide probes [11], and CombiMatrix's (Mukilteo, WA) CustomArray^{TM} (50 - 70mers) which contains arrays of individually addressable microelectrodes for *in situ* oligonucleotide synthesis by means of an electrochemical reaction [12,13].

Commercial arrays are expensive, relatively difficult to customize probe design, and often limited to the model organisms. Spotted long oligonucleotide microarrays provide an inexpensive and highly customizable alternative. These arrays are produced in a similar fashion as the spotted cDNA arrays by depositing solutions of pre-synthesized oligonucleotide probes on a glass slide. The long oligonucleotide probes, usually 40 to 70mer in length, can be synthesized commercially. The array production can normally be performed in an in-house academic facility used for producing cDNA arrays, making it an ideal platform for academic laboratories.

Although spotted oligonucleotide arrays can be produced and used with a very similar method to those widely used for cDNA arrays, the success of oligonucleotide-based arrays are highly dependent on their probe design. To fulfill the objective of an oligonucleotide-based genome array, several design considerations need to be addressed. Most importantly, the probe sequence should be unique in the genome to minimize cross-hybridization. In addition, based on empirical rules used in primer designs, sequences that can form internal secondary structures should be avoided to maximize probe accessibility. Low complexity sequences should also be avoided to prevent nonspecific hybridization [14-16]. Other criteria are more unique to the design of a genome array, such as uniformity in probe melting temperatures and the proximity of probes to the 3' end of a gene. Another critical consideration is the choice of probe length, a balance between specificity and synthesis feasibility. In general, longer probes provide better specificity, but are associated with increasingly lower percentages of full-length probes (assuming 99% coupling efficiency, less than 50% of 70mer probes are full-length) and higher cost. Very short probes (<25mer) such as those used by GeneChip arrays require multiple probes per gene to improve signal specificity.

A computational approach is ideal to find the optimum design solution for this multi-parameter problem. Existing primer design programs are inadequate for designing a whole genome array. Therefore, we developed ArrayOligoSelector (AOS) specifically for the purpose of systematically selecting gene-specific long oligonucleotide probes for entire genomes. For each open reading frame (ORF), the program optimizes the oligonucleotide selection on the basis of several parameters, including uniqueness in the genome, sequence complexity, lack of self-binding, and GC content. Using AOS, we designed a long oligonucleotide microarray for the complete genome of *Plasmodium falciparum*, the most deadly causative agent of human malaria. This malaria chip has been used to study the transcriptome of the parasitic intraerythrocytic developmental cycle and sequence variation between different *P. falciparum* strains.

Similar approaches to oligonucleotide design have previously been described, but the exact algorithms, source code, and/or accompanying hybridization data are not available [8,10,17].

We made the algorithm, AOS source code and software, as well as the hybridization data publicly available to ensure public usage of the program, which is especially important for designing genome arrays for organisms like *Plasmodium* that hold minimal commercial interest, yet are immensely important for public health. Since we made AOS available, scientists all over the world have used AOS to design genome arrays for a wide variety of organisms including mouse, malaria, yeast and bacteria.

Algorithms

To design an optimum set of oligonucleotide probes for a given organism, AOS uses the ORF sequences and the complete genomic sequence as inputs, and then selects an optimum oligonucleotide for each ORF. The workflow of AOS consists of four major steps: 1) data preprocessing to ensure the correct sequence format and user inputs; 2) cognate sequence identification to discriminate true genomic targets from regions of potential cross-hybridization; 3) computing the following parameters for every oligonucleotide in an ORF sequence: uniqueness in the genome, internal secondary structures, GC percentage, and sequence complexity; 4) selecting a set of optimum oligonucleotide sequences using a rule-based filter procedure.

Step I: Data preprocessing

Correct data format, user inputs and computational resource are critical to ensure a smooth AOS execution. In the data preprocessing stage, AOS interacts with a user to obtain the sequence files (ORF sequences and the complete genome sequence), the oligonucleotide probe length, the choice for sequence masking, and the method to identify cognate sequences. It then verifies that the sequences are in the correct FASTA format, sequence identifiers do not contain white space characters to interfere with result parsing, no duplicated sequences in the sequence files, user input parameters are in the correct numerical range and selections, and the appropriate operating system is used. If all checks are passed, AOS is recompiled on the user's computer and proceeds to the next step.

Step II: Cognate genomic sequence identification

The cognate region is the genomic region where an ORF originates. Accurate identification of cognate regions is essential for differentiating true targets for an oligonucleotide from cross-hybridization regions. Since this information may not be easily available to all users, AOS opts to derive this information computationally based on the sequences provided in the two input files.

As the second step in the program workflow, AOS identifies an ORF's cognate region by reconstructing its exon structure. Each ORF was first aligned to the genomic sequences by a sequence homology search (BLAST or BLAT), alignments with 100% identical matches were stitched back together through a heuristic process to recapitulate the exon pattern. The principle behind this strategy is that individual exons should be among those perfect alignments, and the goal is to identify those specific perfect alignments and the exact order they should be arranged in to form the corresponding ORF. However, the difficulty comes from the fact that not every 100% identical alignment region is necessarily a part of the ORF exon structure. Although an exhaustive search of all possible arrangements of any number of perfect alignment regions can find the correct exon structure, the number of arrangement combinations increases factorially as the number of perfect alignments increases($\sum_{n} n!$), which makes an exhaustive strategy impossible to complete if a great number of perfect alignments was initially identified.

Therefore a heuristic approach is used to decrease the search space. The first heuristic trick is that a search can only start from either a perfect alignment of >50 bp, or a "must-use" alignment (see below for details). Secondly, a search can only continue by adding other perfect alignment regions that satisfy the following spatial constraints: same chromosome and strand orientation; proximity to all existing alignment regions (<3000 kb); minimum overlap with existing alignment segments (<70% of the smaller of the new and existing regions); consistent arrangement in both ORF and genome sequences (e.g. if the new region was to the 5' end of an existing region in the ORF sequene, it must be so in the genomic sequence as well). Third, a seach stops when the sum of the existing regions reaches the length of the ORF. Fourth, a search also stops when the sum of all potential regions is highly unlikely to reach the length of the ORF. Fifth, only 100% perfect alignment regions can be considered (SNPs not allowed). Sixth, if the first high scoring hit alignment (best alignment) to any chromosome is less than 50bp, alignment regions from the entire chromosome will not be considered.

In simple terms, the AOS search procedure is to construct combinations of alignment regions; each combination a solution for the correct exon pattern. Procedurely, the above heuristics is implemented as first identifying all perfect alignments, followed by building a connectivity matrix to specify compatible alignments if an alignment has already been selected as part of an exon pattern (based on the spatial constraints described above: compatible chromosome, strand direction, proximity, overlap, and spatial orientation). Subsequently, AOS identifies the “must-use” alignments by scanning for regions in the ORF sequence that are covered by a single perfect alignment, and the corresponding alignment is referred to as the “must-use” alignment. After that, the AOS search starts to construct a list with a single alignment that is either a “must-use” alignment or a perfect alignment >50 bp. AOS proceeds to add additional perfect alignments (n) that are allowed by the connectivity matrix. The original list is duplicated n times and a different alignment is added at the end of each list. This duplication and extension procedure continues until when existing alignments in a list have reached the full length of the ORF. If existing alignments in a list plus all their potential additions (allowed by the connectivity matrix) cannot reach the full length of the ORF, the list is eliminated from furthur consideration. At the end of the search process AOS finds a collection of lists; each contains one or more alignments. Each list is a possible solution for the real exon pattern.

To ensure the accuracy of the results, lists in the final collection are re-examined. Only combinations within ±20 bp of the ORF full length size and able to generate the original ORF sequence in a correct order are kept as solutions for the exon pattern reconstruction. Multiple solutions are allowed. The corresponding exon locations in the genomic sequences are extracted as an ORF's cognate region. This cognate region information is stored in disk to be used in the uniqueness calculation in Step III.

Users can choose to use either the BLAST or BLAT program for sequence alignment to identify the perfect alignment regions [18]. BLAST is more sensitive and typically generates a greater number of alignments, therefore resulting in a bigger search space and slower speed for exon pattern reconstruction. Using BLAT is faster, but it has the risk of missing short alignments. It is important to note that the low complexity filter must be turned off during alignment at this step, otherwise cognate regions will fail to be identified. However, this is at a great cost of computational speed due to the large number of short low complexity alignments generated.

Step III: Parameter computation

In the parameter computation step, AOS calculates values for the following features for every oligonucleotide sequence: uniqueness in the genome, internal secondary structure, sequence complexity, GC percentage, and position in the ORF. Each feature is computed using an independent module, which can also be used as a stand-alone program to obtain individual parameter. The parameter values were written to disk for use in the later selection step.

1. Uniqueness in genome

The uniqueness of an oligo in the genome was measured as the theoretical binding energy of the worst potential cross-hybridization to its homologous regions in the genome. Potential cross-hybridizations are detected by BLASTN alignment, followed by binding energy calculation using the energy module. The uniqueness score of an oligonucleotide is the most stable binding energy between the oligo and the genome excluding the corresponding ORF's cognate region.

In earlier versions of AOS, we used the number of sequence identity in BLAST alignment between the oligo sequence and the genomic cross-hybridization targets as our measurement of cross-hybridization. But our experimental results demonstrated that this metric was a poor predictor for cross-hybridization of different hybridization binding structures.   A DNA-DNA duplex becomes less stable when bulges (sequence mismatches) are introduced into the middle of the duplex. Given the same number of perfect base pairing (sequence identity), hybridization signal strength is stronger when the matches form a continuous stretch compared to a different duplex structure with mismatches distributed in the middle (Figure 3-4).

To overcome the difficulty to predict cross-hybridization by simple sequence identities, we implemented the energy module to calculate hybridization binding energy, in order to unify predictions of different binding structures into a single formulation. The binding energy calculation is based on the nearest neighbor model for calculating nucleic acid helix formation and melting temperatures [19], RNA secondary structure prediction algorithms [20,21], and experimentally estimated thermodynamic free energy parameters for oligonucleotide duplexes and RNA secondary structures [22-28].

In addition to careful modeling of the duplex energetic property, the accuracy of the binding energy calculation is highly dependent on initial accurate identification of those DNA duplexes. AOS uses BLASTN alignment program to identify those regions between the ORF and the genome, and then uses the energy module to calculate the binding energy between the aligned regions.

1.1 Binding energy score

The binding energy score is the summation of the following three terms: the base pair stacking energy between the two adjacent base pairs (such as dAA/dTT), the initial binding energy required for helix initiation, the interior and bulge loop destabilizing energy.

The base pair stacking energy is derived based on the nearest neighbor rules, i.e., the energy of the duplex is the addition of free energy terms of each adjacent Watson-Crick base pair, which includes energy contribution for both base pair stacking and hydrogen bonding. For example, in the following five base pair duplexes, the first two base pairs (dAT/dAT) have a stacking energy of –0.9 kcal/mol, the second and third base pairs (dTT/dAA), –1.2 kcal/mol; the third and fourth base pair (dTG/dCA), –1.5 kcal/mol and so on. The final stacking energy term is (–0.9) + (–1.5) + (–1.2) + (-2.3) = -5.9 kcal/mol.

```
A T T G C
| | | | |
T A A C G
```

Individual stacking energy parameters were obtained by experimentally estimating nearest neighbor parameters for all ten adjacent base pair combinations [22].

The helix initiation energy term models the free-energy change for initiation of DNA duplex, which was estimated experimentally to be +3.4 kcal/mol [22,28].

The interior loop or bulge loop can form when mismatches are closed by at least 2 base pairs. Mismatches on both strands result in the formation of an interior loop. If a mismatch only exists on one strand, the formation is an interior bulge. Both interior loop and bulge contributed destabilizing free energy to the duplex. The loop or bulge destabilizing energy is modeled as the sum of the following three terms: an entropic term that depends on the size of the loop or bulge; terminal stacking energy for the mismatch base pairs adjacent to both closing base pairs, which sometimes provides a favorable free energy; an asymmetric loop penalty for non-symmetric interior loops [20]. The terminal mismatch stacking energy parameters such as dAA/dTA (+0.61 kcal/mol) were estimated experimentally using short nucleic acid duplexes [23-27]. The parameters for the entropic term were derived from parameters used in RNA secondary structure prediction, which were empirical approximations of experimental measurements (Table 2-1) [21].

The parameter for asymmetric loop penalty was based on a study of internal loops in oligonucleotides by Peritz et al. [29,30]. An asymmetric internal loop with a size of $N1$ and $N2$ nucleotides should be penalized by $N* f(M)$ kcal/mol, where $N= | N1 – N2 |$, M is the minimum of 5, $N1$ or $N2$, and $f(1) = 0.7$, $f(2) = 0.6$, $f(3) = 0.4$, $f(4) = 0.2$ and $f(5) = 0.1$.

The nearest neighbor model had good agreement with experimental data on short duplexes. It is well known that the binding free energy and melting temperature of double-stranded DNA molecules plateau at a longer length. However, evidence for size limitation of the nearest neighbor model and parameters is sparse. In addition, the above thermodynamic parameters used in our binding energy calculation were estimated from experimental measurements on short oligonucleotide duplexes (<20 bp). Therefore, although we used both to model long oligonucleotide duplex binding stability, the binding energy values should be viewed as a function of binding stability on a relative scale, rather than be interpreted as the absolute free energy generated during DNA duplex formation.

1.2 Energy score correlates linearly with measured hybridization strength on 70mer oligonucleotides

Although the energy module and parameters are probably not an accurate depiction of the true binding energetic property of long oligonucleotide DNA duplexes, we were interested in using the energy score as a relative measurement of hybridization strength, which could then be used to estimate potential cross-hybridization. To evaluate the utility of the binding energy score to measure cross-hybridization, we conducted experiments on a series of 70mer oligonucleotides with various predicted duplex structures.

We designed several series of 70mer microarray probes that target the *Plasmodium falciparum* genome. In each series, there was a perfect 70mer that matched the coding sequence of an ORF perfectly; the rest of the series was composed of 70mers with various numbers of mismatched base pairs distributed either at the terminals or in the middle of the 70mer. We hybridized transcripts extracted from various stages of *P. falciparum* parasites and then obtained the relative hybridization signal of the mismatched 70mers to the perfect 70mer in each series. The binding energy score of each mismatched 70mer was computed for the duplex (alignment between the mismatched and perfect 70mers). Results demonstrated that there existed a linear relationship (Pearson correlation coefficient *r* = -0.91) between binding energy scores and the relative hybridization strength (Figure 3-3).

1.3 Speed Optimization

Binding energy scores are calculated as the sum of many independent terms, such as the base pair stacking energy and loop destabilizing penalties. Therefore, for two adjacent oligonucleotide probes (with a single base pair offset), their energy score calculation involves a large degree of redundancy. In addition, potential cross-hybridization regions were initially identified by BLAST, followed by the binding score calculation, if we simply used a single oligonucleotide sequence as the input to the energy module, essentially the same BLAST operation would be carried out for adjacent oligonucleotides as well. Both kinds of redundancy would dramatically decrease the speed of the energy module.

To increase the speed, we optimized the energy module by the following strategies. First, we only performed a single BLAST alignment using the entire ORF sequence. Second, we computed the free binding energy score for an entire alignment instead of for a single oligonucleotide, excluding any alignment from the cognate sequence region. Third, in addition to a single energy score, we recorded the score contributions from every adjacent base-pair in the entire alignment. To derive the binding energy score for an oligonucleotide, we simply summed up the score contributions from the corresponding regions in the alignment. Since an oligonucleotide sequence could be covered by more than one alignment, the final binding energy score was the most stable energy score (the largest absolute value).

2. **Internal secondary structure**

The secondary structure module measures the potential of forming an internal hairpin structure within an oligonucleotide. A fast approximation for detecting internal hairpins is by aligning the oligo sequence with its reverse compliment. We implemented the Smith-Waterman algorithm to search for the optimal local alignment [31] and used the alignment score to represent the potential to form internal hairpins. PAM47 DNA matrix is used (match +5, mismatch –4, gap opening –7, gap extension 0) in the implementation for local alignment.

Sophisticated RNA secondary structure prediction methods such as Mfold were available [32] and likely to generate more accurate results, but they are much slower computationally.

3. **Sequence complexity**

The sequence complexity module measured the level of oligo sequence complexity using the LZW compression algorithm [33]. The advantages of this method are fast computational speed and no need for prior information for low complexity sequence elements. It is implemented as the size of the oligonucleotide sequence minus its compressed version in bytes.

4. **GC content**

GC content is a key factor determining DNA duplex melting temperature. We used it as the proxy for melting temperature, calculated as the the number of G C base-pairs over the length of the oligo.

Step IV: Optimum selection

The last step of the AOS algorithm is to select a set of optimum oligonucleotide sequences based on the parameters computed in Step III. An ideal oligo probe has a small negative value of binding energy score (unique in the genome), a small secondary structure score (lack of internal hairpins), a small sequence complexity score, a %GC close to the user-defined target %GC, and close to the 3'end of the ORF sequence.

We implemented a rule-based filtering procedure to select for the optimum oligonucleotide. The first filter is the uniqueness filter. Oligos belonging to a single ORF are ranked first by their uniqueness scores (binding energy score). Oligos scoring better than both an optional user-defined threshold and the default cutoff are kept in the candidate pool. The default cutoff is defined as the larger (smaller absolute value, note energy scores were negative values) of the following two terms: the 5th percentile in the rank, and the best uniqueness score minus 5 kcal/mol.

The second filter is to eliminate any oligos with user-defined (optional) unwanted sequences, such as a long stretch of AT sequence.

The third filter operates on the sequence complexity parameter and secondary structure score in parallel. Similar to the operation on energy scores, oligos that pass the cutoffs can proceed further. Although it only operates on the current candidate pool (oligos that passed the previous two filters), the cutoffs are determined using the complete set of oligos belonging to a single ORF. The initially cutoffs are determined at the top 33rd percentile of the rank by either the secondary structure score or the sequence complexity score. If there is no candidate oligo that can pass both thresholds simultaneously, each cutoff is relaxed incrementally (secondary structure score cutoff increases by 10; sequence complexity score cutoff increases by 1) until one or more oligos pass both thresholds simultaneously.

The fourth filter operates on the %GC parameter. Initially, only oligos with the user-defined target %GC can pass. If no oligo in the current candidate pool satisfies this criterion, the %GC boundaries are relaxed by 1 percentage point at a time in each direction until one or more oligos score within the range.

The final filter operates on the 3' proximity to select the oligo that is closest to the 3' end of the parent ORF. This oligonucleotide is our optimum selection. At this point, AOS reaches its final step to generate program output of the optimum oligo selection.

Occasionally, if a user wants to design more than one oligo per ORF, AOS will attempt to select non-overlapping (must be >10 bp apart at the oligo starting positions, but typically >50bp) oligos from the current pool. If this is not successful using the current candidates, the selection procedure iterates from the combined secondary structure and sequence complexity filter to the 3' proximity filter, until the desired number of oligos is selected, or when the cutoffs are fully relaxed and the candidate pool reaches its maximum size.

Reference

1. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686.

2. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional discovery via a compendium of expression profiles. Cell 102: 109-126.

3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511.

4. Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, et al. (2005) The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat Genet 37: 544-548.

5. Rubins KH, Hensley LE, Jahrling PB, Whitney AR, Geisbert TW, et al. (2004) The host response to smallpox: analysis of the gene expression program in peripheral blood cells in a nonhuman primate model. Proc Natl Acad Sci U S A 101: 15190-15195.

6. Hardiman G (2003) Microarrays Methods and Applications; Hardiman G, editor. Eagleville: Dna Press.

7. Hardiman G (2004) Microarray platforms--comparisons and contrasts. Pharmacogenomics 5: 487-502.

8. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14: 1675-1680.

9. Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, et al. (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res 12: 1749-1755.

10. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, et al. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol 19: 342-347.

11. Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, et al. (2002) An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. Nucleic Acids Res 30: e30.

12. Tian J, Maurer K, Tesfu E, Moeller KD (2005) Building addressable libraries: the use of electrochemistry for spatially isolating a heck reaction on a chip. J Am Chem Soc 127: 1392-1393.

13. Tesfu E, Maurer K, Ragsdale SR, Moeller KD (2004) Building addressable libraries: the use of electrochemistry for generating reactive Pd(II) reagents at preselected sites on a chip. J Am Chem Soc 126: 6212-6213.

14. Chavali S, Mahajan A, Tabassum R, Maiti S, Bharadwaj D (2005) Oligonucleotide properties determination and primer designing: a critical examination of predictions. Bioinformatics 21: 3918-3925.

15. van Baren MJ, Heutink P (2004) The PCR suite. Bioinformatics 20: 591-593.

16. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

17. Rouillard JM, Herbert CJ, Zuker M (2002) OligoArray: genome-scale oligonucleotide design for microarrays. Bioinformatics 18: 486-487.

18. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.

19. Tinoco I, Jr., Uhlenbeck OC, Levine MD (1971) Estimation of secondary structure in ribonucleic acids. Nature 230: 362-367.

20. Lyngso RB, Zuker M, Pedersen CN (1999) Fast evaluation of internal loops in RNA secondary structure prediction. Bioinformatics 15: 440-445.

21. Turner DH, Zuker M (2005) Free Energy and Enthalpy Tables for RNA Folding.

22. Sugimoto N, Nakano S, Yoneyama M, Honda K (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. Nucleic Acids Res 24: 4501-4505.

23. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J, Jr. (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. Biochemistry 38: 3468-3477.

24. Allawi HT, SantaLucia J, Jr. (1998) Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. Biochemistry 37: 9435-9444.

25. Allawi HT, SantaLucia J, Jr. (1998) Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. Biochemistry 37: 2170-2179.

26. Allawi HT, SantaLucia J, Jr. (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. Biochemistry 36: 10581-10594.

27. Allawi HT, SantaLucia J, Jr. (1998) Thermodynamics of internal C.T mismatches in DNA. Nucleic Acids Res 26: 2694-2701.