

Simulated Sequencing by Hybridization Using Evolutionary Programming

Gary B. Fogel

Natural Selection, Inc.
3333 N. Torrey Pines Ct., Suite 200
La Jolla, CA
USA 92037
gfogel@natural-selection.com

Kumar Chellapilla

Dept. of ECE
U. C. San Diego
La Jolla, CA
USA 92093-4007
kchellap@ece.ucsd.edu

Abstract

Sequencing of DNA is among the most important tasks in molecular biology. DNA chips are considered to be a more rapid alternative to more common gel-based methods of sequencing. Previously, we demonstrated the reconstruction of DNA sequence information from a simulated DNA chip using evolutionary programming. The research presented here extends this work by relaxing several assumptions adopted in our initial investigation. We also examine the relationship between base composition of the target sequence and the useful set of probes required to decipher the target on a DNA chip. Comments regarding the nature of the optimal ratio for the target and probe lengths are offered. Our results go further to suggest that evolutionary computation is well-suited to address the sequence reconstruction problem.

1. Introduction

DNA chips are becoming increasingly useful tools for gene expression analysis (deSaizieu et al., 1998; Lipshutz et al., 1999; Duggan et al., 1999), identification of single-nucleotide polymorphisms in genes (Wang et al., 1998), mapping of allelic variation in genomes (Winzeler et al., 1998), and rapid sequencing of DNA (Cantor et al., 1992; Pease et al., 1994). A solution containing a DNA fragment to be sequenced is exposed to an array (referred to as a *DNA chip*) containing all possible DNA fragments of a pre-specified smaller length (referred to as *probes*). The probes on the chips are built by combinatorial chemical synthesis and are chemically attached to the substrate (Ross, 1996). In some cases, these probes will bind or “hybridize” with the target sequence due to complementary base pairing rules. Probes with no base complementation to the target will not hybridize. The original target sequence can then be reconstructed using overlapping probe sequences and knowledge of the position and sequences of the hybridized probes (Figure 1). This mapping is usually performed with the use of fluorescent tags. A laser scans the chip, excites the fluorescent tags, and a computer records the fluorescence pattern (Gibbs, 1996). The fluorescence across all positions is referred to as the *fluorescence pattern (FP)*. The entire process is known as sequencing by hybridization (SBH) (Cantor et al., 1992). DNA chips commonly are made with the set of all possible probes eight nucleotides in length (referred to as *octamers*) generating 65,536 unique probes spaced on a 1.6 cm² array (Fodor et al., 1991). Similar chips have already been used to identify mutations in HIV associated with drug resistance (Lipschutz, 1995) and for identification of mutations in a gene associated with breast cancer (Hacia et al., 1996). The problem at hand consists of correctly reconstructing the sequence of a *target* DNA string of nucleotides given the number of symbols in the DNA sequence, N , and the associated fluorescence pattern, FP_{ref} , generated by washing a DNA chip with fluorescently labeled DNA fragments. Each of the grid positions in the DNA chip contains a unique probe of length n .

For ease of description, consider the example DNA target sequence 5'-ATTGATTCG-3', with length $N = 9$ and a DNA chip with all possible probes of length $n = 4$. A DNA chip with probe length n will have 4^n positions in the grid on the DNA chip (and a similarly sized *FP* space). So, for a probe length of 4 there exist 256 grid positions, each associated with a unique probe sequence. All possible 4-nucleotide probes would exist in the set: {AAAA, AAAC, AAAG, AAAT, AACA, ..., TTTA, TTTC, TTTG, and TTTT}.

When a solution of target DNA is washed over the DNA chip, the probes in each of the grid positions have a potential to bind to the target DNA. On binding, the position will fluoresce when illuminated by a laser, generating a *FP* that can be visually observed. For the results presented here, the light intensity in each position is considered to be proportional to the maximum number of bindings that occurs between the target and probe sequences. For example, when the target 5'-ATTGATTCG-3' is washed over a grid position with the probe 3'-TAAC-5', the probe would bind to complementary sequences 5'-ATTG-3', 5'-TTGA-3', 5'-TGAT-3', 5'-GATT-3', 5'-ATTC-3', and 5'-TTCG-3' in the target with binding “efficiencies” of 4, 1, 0, 1, 3, and 2, respectively. The corresponding entry in the *FP* would be 4, i.e., max(4,1,0,1,3,2). It is well known that there is significant variation in duplex stability between GC pairings

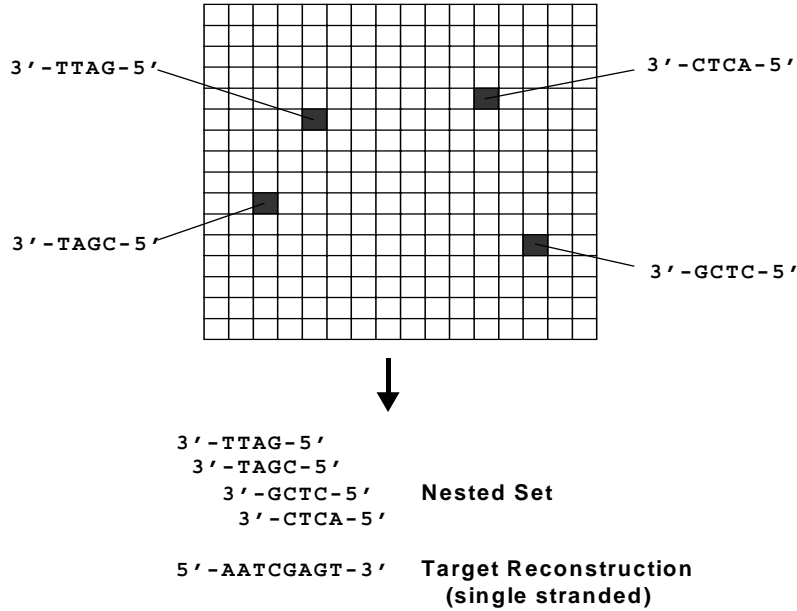


Figure 1. Sequencing by hybridization method. A DNA chip of all possible probes of length n (in this case 4) is washed with a target sequence of length N (in this case 8). By knowing the sequence of the probes in each of the grid positions, it is possible to generate the target sequence as both single and double stranded due to base complementation.

and AT pairings, however several new strategies exist in the literature to obtain DNA duplexes with thermal stability independent of their AT/GC ratio content (Nguyen et al., 1997; Hacia et al., 1998). For the results presented here, the thermal stability of GC and AT pairings are considered equal.

Previously, we developed a simulation of the sequencing by hybridization process and used evolutionary programming (Fogel *et al.*, 1966, Fogel, 1995) to determine the most suitable target lengths for the set of all tetramer probes (Fogel et al., 1998). In these experiments, all entries in the FP that had a value below $(n/2)$ were set to 0. With a probe length of n , all of the elements in the FP space had values in $\{0, n/2, n/2+1, \dots, n\}$. For the results presented here, this restriction has been relaxed such that all possible FP values are now used. Hybridization sensitivity is increased and is more closely correlated to established fluorescence detection systems reported in the literature. For instance, commercially available fluorescence detection systems are capable of distinguishing a single base mismatch in a 25-mer oligonucleotide.

Our previous research was limited to a probe length of 4 and target lengths varied from 10 to 40. In the current study, the probe lengths are extended to include 5 and 6, and the target lengths range from 10 to 100. This has generated a clearer understanding of optimal target and probe length ratios. Repetition in the target sequence appears to be a limiting factor in the usefulness of a DNA chip in sequencing. We present alternative methods that could be used to avoid this problem in future simulations.

2. Methods

The target sequence consisted of the DNA subsequence of length N that was randomly generated in advance and fixed during the simulation to assess the quality of evolved solutions. Let the fluorescence pattern generated by the true target sequence be denoted by FP_{ref} . The error of any given test solution sequence of length N was computed using these steps:

1. Compute the fluorescence pattern, FP_{test} , of the test sequence for the given probe length.
2. FP_{test} was compared with FP_{ref} and the sum of the absolute differences over all the entries in the FP was computed.
3. This sum constituted the FP error score of the test solution. The normalized FP score was obtained by dividing the FP score by the number of non-zero entries in FP_{ref} . A test solution that perfectly matched the actual

solution had an error of 0.

Evolutionary programming was used to search for the optimal test sequence with the lowest *FP* error score using the following steps:

1. *Initialization*: A population of M initial parent solution strings, $\{S_1, S_2, \dots, S_M\}$, was generated. Each initial parent string, S_i , was produced by randomly selecting (with uniform probability) N symbols with replacement from $\{A, C, G, T\}$. These solutions constituted the parents at generation zero. Every S_i had three strategy parameters, p_{1i} , p_{2i} , and p_{3i} , associated with it that were set (somewhat arbitrarily) to four. The generation number, g , was set to one.
2. *Mutation*: Each parent solution, S_i , was mutated to generate one offspring solution through the application of one of three mutation operators, namely, *PointMutate*, *RotateMutate*, and *InsertMutate*. The mutation operator to be applied was selected at random over the three operators with equal probability. The offspring's strategy parameter, p'_{ki} , corresponding to the selected mutation operator (represented by k) was obtained using

$$p'_{ki} = p_{ki} + 0.5N(0,1) \quad (1)$$

while the remaining strategy parameters were copied from the parent. If p'_{ki} exceeded N or fell below 1 it was set to the limit it violated. *PointMutate* randomly selected J_1 symbols in the parent and replaced them with randomly selected symbols from $\{A, C, G, T\}$. J_1 was obtained by sampling a Poisson random variable with mean p'_{1i} . If J_1 exceeded N or fell below 1, it was reset to the limit it violated. *RotateMutate* randomly rotated the parent sequence cyclically by J_2 symbols. The direction of cyclic rotation, clockwise or counterclockwise, was selected uniformly at random. J_2 was obtained by sampling a Poisson random variable with mean p'_{2i} . If J_2 exceeded $N/2$ or fell below 1, it was reset to the limit it violated. *InsertMutate* randomly selected a contiguous subsequence of length J_3 in the parent string, removed it and reinserted it at a randomly selected location in the remaining string. J_3 was obtained by sampling a Poisson random variable with mean p'_{3i} . If J_3 exceeded $N/2$ or fell below one, it was reset to the limit it violated.

3. *Fitness evaluation*: Each member in the population was evaluated in light of the normalized *FP* error score.
4. *Tournament*: Each member was compared with 10 opponents that were randomly selected (with replacement) from the population. The member received a "win" for each comparison in which the error of the member was lower than or equal to that of the opponent.
5. *Selection*: The M members with the highest number of wins were selected to be the parents for the next generation. The generation number, g , was incremented.
6. Steps 2 to 5 were repeated for a predefined number of generations, g_{max} .

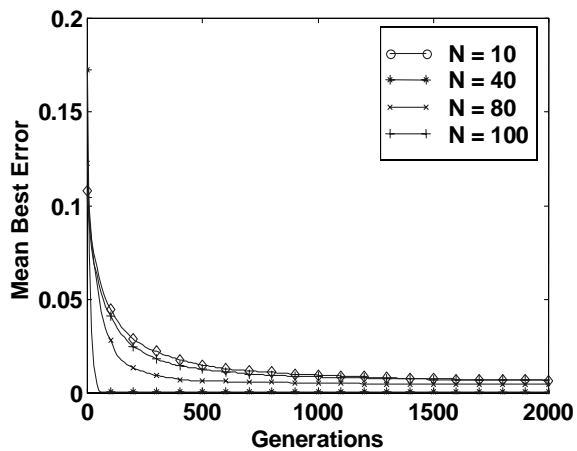
3. Simulations and Results

Experiments were conducted to test the viability and efficiency of the evolutionary programming (EP) procedure. In all the experiments, 100 parents were selected for evolution, the maximum number of generations, g_{max} , was set to 2000, resulting in a total of 200,000 function evaluations during each trial. The target DNA sequence length was chosen from $\{10, 20, 30, 40, 50, 80, 100\}$. Fifty independent trials were conducted for probes of length 4, 40 independent trials for probes of length 5, and 20 independent trials for probes of length 6. The average rate of error optimization of EP for probes of length four, five, and six against all target lengths are shown in Figure 2 give the corresponding rate of optimization of the normalized error, obtained by dividing the *FP* error scores by the sum total of all non-zero entries in the *FP* used during each of the trials. Across all probes, as the ratio of target to probe length increased, the rate of optimization decreased and EP took longer to find the optimal solution.

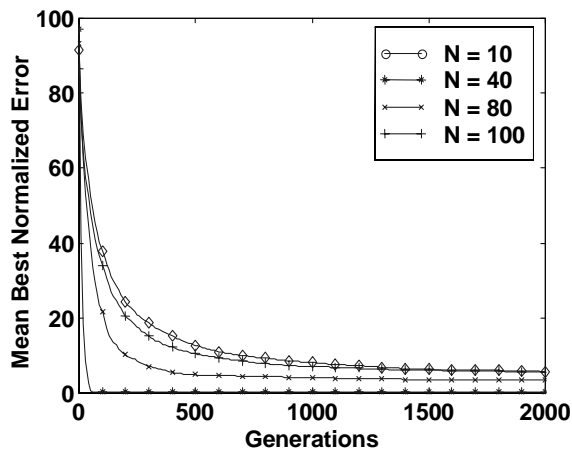
All trials were successful (i.e., zero error) for probes of length five and six and target sequences of length 10. Figure 3 depicts the mean best normalized error value at the end of each trial. Lower error values indicate higher quality solutions. The percentage accuracy was determined as the fraction of successful trials (zero error) for each target and probe combination (Figure 4a). As expected, percentage accuracy decreased as target length increased for any given probe length.

The probability of success for any probe and target combination can be increased by performing multiple trials with different initial populations. Figure 4(b,c) depicts the number of independent trials, R , that need to be conducted with a population size of 100, lasting 2000 generations, to achieve a 90% probability of success. The R values were computed using

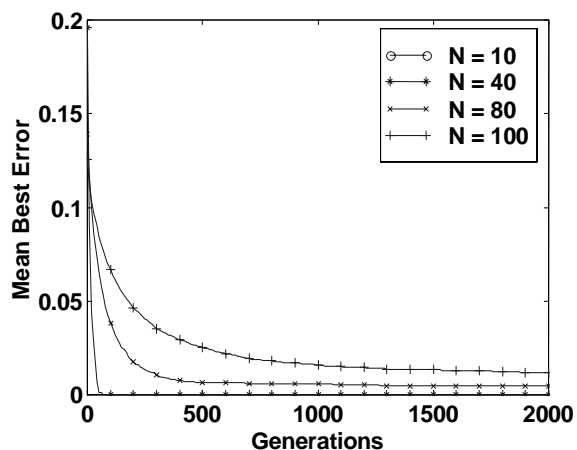
$$R = \text{ceil}\left(\frac{\log(1-z)}{\log(1-p_s)}\right) \quad (2)$$



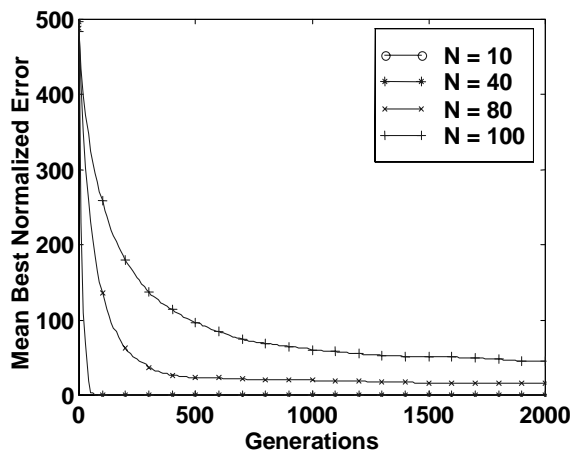
(a)



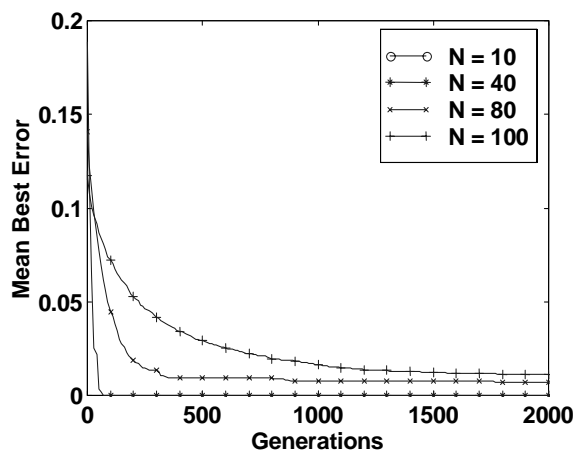
(b)



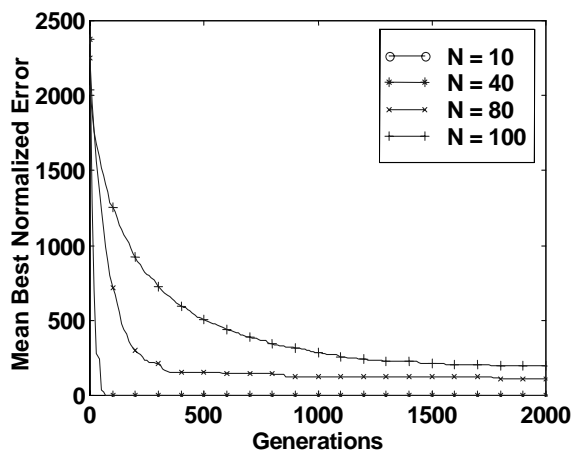
(c)



(d)



(e)



(f)

Figure 2. The mean best error and mean best normalized error minimization averaged over multiple trials following EP. Probe lengths of four (Fig. 2a and 2b), five (Fig. 2c and 2d), and six (Fig. 2e and 2f) were used. Target lengths (N) increase from the left to the right in each figure as 10, 40, 80, and 100 nucleotides. The normalized error values were computed by dividing the error values by the sum total of all the non-zero entries in the fluorescence pattern of the target sequence used during the trial.

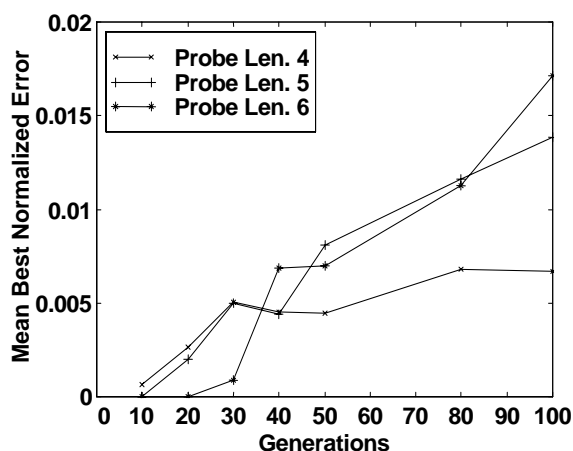


Figure 3. The mean best normalized error values at the end of the trial for target sequence lengths in {10, 20,..., 100}. The population size was 100, the probe lengths were 4, 5, and 6 and each trial lasted 2000 generations. The standard error values were smaller than the size of the line markers at each point and have been omitted for clarity.

where $z = 0.90$ and p_s is the fraction of successful trials. This equation has been corrected from a typographical error in our previous publication (Fogel et al., 1998). As the target length increased from 10 to 100 nucleotides, the quality of the final solution gradually decreased with the lowest quality occurring at a target length of 40, 80, and 100 for probes of length 4, 5, and 6 respectively.

4. Discussion

Previous results demonstrated that evolutionary programming was well suited to SBH (Fogel et al., 1998). In SBH, an appropriate length probe must be used to unambiguously determine a target of length N . When N is large (over 40 nucleotides), a probe of length 4 cannot be used to reconstruct the target sequence with a high probability of success (Fogel et al., 1998). As N increases, the probability of redundancy in the target increases making unambiguous reconstruction difficult (Noble, 1995). Similarly, small probes are likely to find complements in long target sequences. This will generate a chip that fluoresces with equal intensity at all positions and lacks useful information.

However, a target sequence of length 4 will bind completely to only one position in the grid and will be unambiguously determined (barring mismatch binding at other grid locations). Target sequences of length four are not useful in terms of learning about genomes on the order of 10^9 nucleotides. Therefore, each probe length is useful over a specified range of target lengths. Above and below this range, all targets reach a saturation of signal. For instance, our previous analysis (Fogel et al., 1998) suggested that probe lengths of 4 nucleotides were best suited to target lengths of 25 to 35 nucleotides in agreement with data in the literature (Bains 1991). Our previous simulations using EP for sequence reconstruction utilized a population size of 500 in trials of 1000 generations (Fogel et al., 1998). Here we use a population size of 100 for trials of 2000 generations. Comparison of results suggests that there is no appreciable improvement. The operational complexity of the algorithm is a function of the population size, probe length, and target length. The complexity increases linearly with the population size and target length, while it increases exponentially with increasing probe length.

Figure 4 suggests that probes of length five can be used to successfully reconstruct target sequences of up to length 45 to 50, with a 90 percent probability of success, given that 6 independent trials can be conducted. Above this length, error scores increase and the required number of trials may become impractical. Figure 4 also suggests that probes of length six can successfully reconstruct target sequences of length 50 to 55 with a 90 percent probability of success given 6 independent trials. Our data suggest that the ratio of target to probe length is roughly exponential across probe lengths 4, 5, and 6. Our results suggest that probes of length 8 could be used to unambiguously determine target of length 80, a value that corresponds to other previous simulations (Belyi and Pevzner, 1997).

On one occasion when using probe length 4, a “simple” target of length 10 was not perfectly reconstructed in 2000 generations. Further inspection of this randomly generated target 5’-ACTATAATCT-3’, revealed that there were many repetitions in the sequence. For instance the di-nucleotides “TA”, “AT” and “CT” were repeated twice. The best

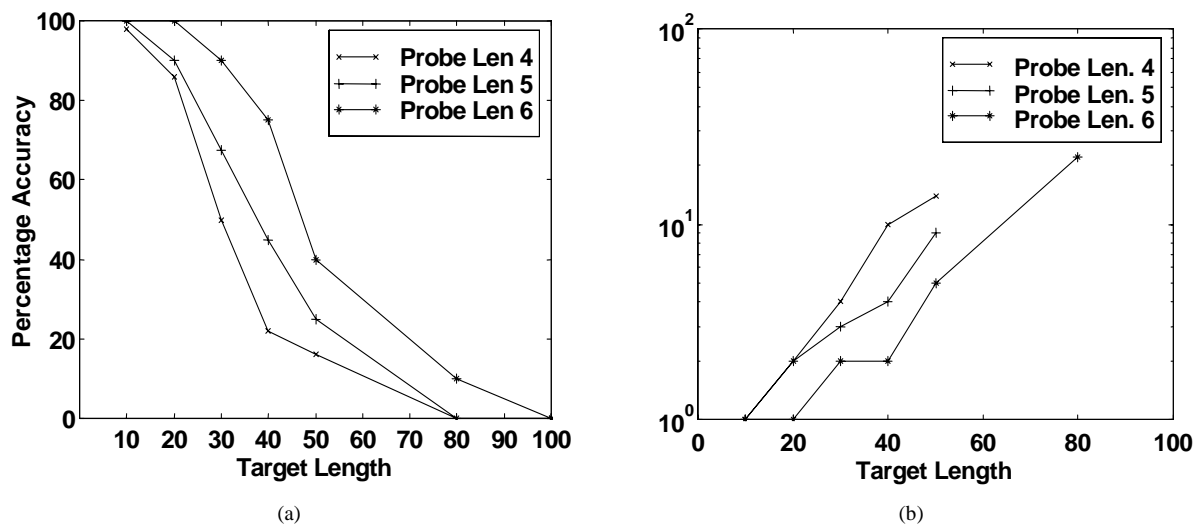


Figure 4. (a) The percentage accuracy in correctly determining a given target sequence with zero error and (b) the number of trials (R) required to achieve a 90% probability of success and target sequence length for probes of length 4, 5, and 6. The number of trials required for 90% probability of success is non-linear and appears to grow exponentially.

reconstructed sequence after 2000 generations was 5'-TAATCTATAA-3'. In essence, repetitions in the sequence precluded the discovery of the optimal solution by allowing sub-optimal solutions to generate significantly high FP scores.

Inspection of trials that did not lead to perfect solutions in other probe and target length combinations lead to the discovery of similar repetitious sequences. These repetitions were either of the type 5'-PPPPPPPPP-3', 5'-PQPQPQPQPQ-3', or 5'-PQQPQQPQQP-3' with R number of unit repeats where P and Q are any different nucleotides from {A,T,C,G} and R is two or more. These repeats appear to be characteristic features of non-perfect trials in our simulations.

Continuous stacking hybridization (CSH) has been used to resolve this problem in real DNA chips (Parinov et al., 1996). CSH employs multiple probe sets in succession. The first probe set acts in the manner previously described. The second probe set is "stacked" on top of the first set and is used to bridge gaps between hybridized probes from the first set. The net effect can theoretically make an octamer chip as efficient as a 13-mer chip when using 5-mer probes in a stacked arrangement (Lysov et al., 1994). This can yield sequencing runs on the order of 1000 nucleotides in length. Parinov et al. (1996) suggested that a similar method utilizing two differently labeled 5-mers stacked to each other and to immobilized octamers can provide sufficient information to reconstruct a DNA sequence of 1000 nucleotides containing repeats up to 16 nucleotides in length. This approach can be incorporated into future experimentation with our EP simulation.

Deterministic methods can be used for reconstructing sequence information from DNA chip fluorescence patterns when there is no noise in the system. To date, our simulations have no noise in the fluorescence signature. However, any noise (as the result of measurement error by the scanning laser or incomplete target hybridization) will significantly degrade the performance of a deterministic algorithm. We believe that evolutionary computation will be quite robust in reconstructing sequences even in the presence of noise in the fluorescence signature. Estimates of noise will be added in future simulations.

References

- Bains, W. (1991) "Hybridization methods for DNA sequencing," *Genomics* 11:294-301.
- Belyi, I. and Pevzner, P.A. (1997) "Software for DNA sequencing by hybridization," *CABIOS* 13(2):205-210.
- Cantor, C.R., Mirzabekov, A. and Southern, E. (1992) "Report on the sequencing by hybridization workshop," *Genomics* 13:1378-1383.
- deSaizieu, A., Certa, U., Warrington, J., Gray, C., Keck, W., and Mous, J. (1998) "Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays," *Nature Biotechnology* 16(1):45-48.

- Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J.M. (1999) "Expression cDNA microarrays," *Nature Genetics* suppl. 21:10-14.
- Fodor, S.P.A., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., and Solas, D. (1991) "Light-directed, spatially addressable parallel chemical synthesis," *Science* 251:767-773.
- Fogel, D.B. (1995) *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, Piscataway, NJ: IEEE Press.
- Fogel, G.B., Chellapilla, K., and Fogel, D.B. (1998) "Reconstruction of DNA Sequence Information from a Simulated DNA Chip Using Evolutionary Programming," In *Proceedings of the Seventh Annual Conference on Evolutionary Programming*, V. W. Porto, N. Saravanan, D. Waagen, and A. E. Eiben (eds.), Springer-Verlag, Berlin pp. 429-436.
- Fogel, L.J., Owens, A.J., and Walsh, M.J. (1966) *Artificial Intelligence Through Simulated Evolution*, Wiley and Sons, Inc. New York.
- Gibbs, W.W. (1996) "New chip off the old block," *Sci. Am.* Sept.:42-44.
- Hacia, J.G., Brody, L.C., Chee, M.S., Fodor, S.P.A., and Collins, F.S. (1996) "Detection of heterozygous mutations in *BRCA1* using high density oligonucleotide arrays and two-color fluorescence analysis," *Nature Genetics* 14:441-447.
- Hacia, J.G., Woski, S.A., Fidanza, J., Edgemon, K., Hunt, N., McGall, G., Fodor, S.P.A., and Collins, F.S. (1998) "Enhanced high density oligonucleotide array-based sequence analysis using modified nucleoside triphosphates," *Nuc. Acids Res.* 26(21):4975-4982.
- Lipshutz, R.J., Morris, D., Chee, M., Hubbell, E., Kozal, M.J., Shah, N., Shen, N., Yang, R., and Fodor, S.P.A. (1995) "Using oligonucleotide probe arrays to access genetic diversity," *BioTechniques* 19(3):442-447.
- Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., and Lockhart, D.J. (1999) "High density synthetic oligonucleotide arrays," *Nature Genetics* suppl. 21:20-24.
- Lysov, Y.P., Chernyi, A.A., Balaeff, A.A., Beattie, K.L. and Mirzabekov, A.D. (1994) *J. Biomol. Struct. Dyn.* 11:797-812.
- Nguyen, H.-K., Auffray, P., Asseline, U., Dupret, D. and Thuong, N.T. (1997) "Modification of DNA duplexes to smooth their thermal stability independently of their base content for DNA sequencing by hybridization," *Nuc. Acids Res.* 25(15):3059-3065.
- Noble, D. (1995), "DNA sequencing on a chip," *Anal. Chem.* 67(5):201A-204A.
- Parinov, S., Barsky, V., Yershov, G., Kirillov, E., Timofeev, E., Belgovskiy, A., and Mirzabekov, A. (1996) "DNA sequencing by hybridization to microchip octa- and decanucleotides extended by stacked pentanucleotides," *Nuc. Acids Res.* 25(15):2998-3004.
- Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P. and Fodor, S.P.A. (1994) "Light-generated oligonucleotide arrays for rapid DNA sequence analysis," *Prod. Natl. Acad. Sci. USA.* 91:5022-5026.
- Ross, D.W. (1996) "DNA on a chip," *Arch. Pathol. Lab. Med.* 120:604-605.
- Wang, D.G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittman, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M., and Lander, E.S. (1998) "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome," *Science* 280:1077-1082.
- Winzler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J., and Davis, R.W. (1998) "Direct allelic variation scanning of the yeast genome," *Science* 281:1194-1197.