Towards DNA Sequencing Chips

Pavel A.Pevzner[1] * and Robert J. Lipshutz[2] **

[1] Department of Computer Science and Engineering
The Pennsylvania State University
University Park, Pennsylvania 16802

[2] Affymetrix, Inc.
Santa Clara, California 95051

Abstract. *DNA sequencing is an important technology for the determination of the sequences of nucleotides that make up a given DNA fragment. In view of the limitations of current sequencing technology, it would be advantageous to have a DNA sequencing method that provides the sequences of long DNA fragments and is amenable to automation. Sequencing by Hybridization (SBH) is a challenging alternative to the classical sequencing methods. The basic approach is to build an array (Sequencing Chip) of short DNA fragments of lenght l and to use biochemical methods for finding all substrings of lenght l of an unknown DNA fragment. Combinatorial algorithms are then used to reconstruct the sequence of the fragment from the l-tuple composition. In this article we review biochemical, mathematical, and technological aspects of SBH and present a new sequencing chip design which might allow significant chip miniaturization without loss of the resolution of the method.*

1 Introduction

DNA sequencing is an important technology for the determination of the sequences of *nucleotides* (referred to as A, C, G, T) that make up a given DNA fragment. Using current sequencing technologies the most proficient laboratories can today sequence 25,000-125,000 nucleotides per person per year at a cost of several dollars per nucleotide. One of the goals set out by the Human Genome Project is to increase the sequencing rate by an order of magnitude and reduce the cost by an order of magnitude. Significant strides have been made in improving existing technologies using automation and incremental improvements to the instruments, however a ten-fold increase in throughput will probably require an entirely new technology.

Sequencing by Hybridization (SBH) is a new approach to DNA sequencing proposed simultaneously and independently by Drmanac and Crkvenjakov,1987 [5], Bains and Smith,1988 [2], Lysov et al,1988 [16], Southern,1988 [24] and Macevicz,1989 [17]. Sequencing by hybridization relies on the following biochemical phenomenon. Given a short (8-30 nucleotides) piece of DNA, called an *oligonucleotide* or a *probe*, and a single-stranded target DNA fragment; the probe will bind (*hybridize*)

* The research was supported in part by the National Science Foundation under the grant CCR-9308567 and by the National Institutes of Health under the grant 1R01 HG00987-01
** The research was supported in part by the National Institutes of Health under the grant HG-00813 and by the Department of Energy under the grant DE-FG03-92-ER81275

to the target if there is a substring of the target that is Watson-Crick *complement* to the probe (A is complementary to T and G is complementary to C). For example a probe $ACCGTGGA$ will hybridize with a target $CCCTGGCACCTA$ since it is complementary to the substring $TGGCACCT$ of the target. In this manner oligonucleotides can be used to 'probe' the unknown target DNA and determine its substring content. Sequencing by hybridization exploits this process. The simplest SBH technique can be described as follows:

- Attach all possible probes of length l (l=8 in the first SBH papers) to the surface of a substrate, each probe at a distinct and known location. This set of oligonucleotides is called the *sequencing chip*.
- Apply a solution containing a radioactively or fluorescently labeled target DNA fragment to the sequencing chip.
- The radioactively or fluorescently labeled single-stranded target DNA fragment hybridizes with those probes that are complementary to substrings of length l of the target fragment.
- Detect oligonucleotides that hybridized with the target fragment with a nuclear or spectroscopic detector. The *oligonucleotide or substring content* of the target DNA fragment is obtained.
- Apply a combinatorial algorithm to reconstruct the sequence of the target DNA fragment from the oligonucleotide content.

During the last 5 years various researchers have been developing SBH and a dozen variations and modifications of SBH have been proposed. Even today a chip for sequencing hundreds to thousands of nucleotides might cost from a few dollars to tens of dollars when made by mass production. The sequencing procedure using such a chip could easily be automated, and the speed of such sequencing on an automated instrument could approach million of bases per day [3].

The current state of the SBH chip technology is described below.

- Mirzabekov's laboratory [14] started a project based on depositing separately synthesized oligonucleotides. A chip was designed where oligonucleotides were immobilized within 100 x 100 micron dots deposited at 100 micron intervals.
- Southern's laboratory built a small sequencing chip containing 4096 oligonucleotides on a large 20 cm x 20 cm glass plate [3, 18]. This group conducted parallel synthesis of oligonucleotides on the glass plate using physical masking. The 4096-oligonucleotide chip has been used to repeatedly demonstrate the feasibility of SBH using surface bound oligonucleotides.
- Drmanac et al. [6] suggested *combinatorial oligonucleotide synthesis* using micro beads.
- Beattie's laboratory [3] suggested a *segmented synthesis* approach currently implemented by Genosys Biotechnologies, Inc. Genosys' current prototype instrument is capable of synthesizing 100 oligonucleotides simultaneously with a cycle time of six minutes per base. If ten of these Genosys machines were put to work, the entire library of 1,048,576 decamers could be prepared in 200 days. However, after synthesis the probes must be attached to the chip.
- Eggers et al. [9] proposed a *genosensor* technology for large-scale oligonucleotide arrays. Genosensors consist of electronically addressable micro-sized dielectric

test fixtures, each containing a synthetic oligonucleotide probe. Microdetection of hybridization is achieved by interrogating the miniature test fixture with a low voltage alternating electric field.

- A most promising approach to high-density chip manufacturing has been developed at Affymetrix. Their method is based upon a newly developed technique for *light-directed polymer synthesis* [10]. Using this technique, building a chip $C(k)$ with all 4^k oligonucleotides of length k requires just $4 \cdot k$ separate reactions. Chips are read using a modified confocal laser microscope (Fodor et al.,1993 [11]). This instrument is a closed system for real time hybridization and analysis of fluorescent intensities. The technique includes a light-directed combinatorial synthesis strategy, a procedure for substrate derivatization, photolabile 5'-protected nucleosides, a lithographic apparatus, and a detection system with detection software (Pease et al.1994 [20]). Huang et al. (personal communication) have recently synthesized a chip with all 65,536 8-tuples, hybridized it to a 16-nucleotide target DNA and reconstructed the target sequence from the hybridization intensity data.

In this paper we describe combinatorial problems related to DNA sequencing chips. We begin (Section 2) by reviewing algorithms for reconstructing DNA sequences from *ideal* hybridization data. In Section 3 we describe several approaches to analyze *real* hybridization data. Problem of optimal chip design is discussed in Section 4. Finally, in Section 5 we design new SBH chips and demonstrate their advantages in comparison with classical chips.

2 Ideal SBH Sequence Reconstruction

Suppose we are given *all* substrings of length l of an unknown string (oligonucleotide *spectrum* of a DNA fragment). How do we reconstruct the target DNA fragment from this data?

2.1 Naive approaches to SBH sequence reconstruction

Oligonucleotides (*probes*) p and q of length l (*l-tuples*) *overlap* if the last $l-1$ letters of p coincide with the first $l-1$ letters of q. Given the spectrum S of a DNA fragment construct the directed graph $H(S, E)$ with vertex set S and edge set $E = \{(p, q) : p \text{ and } q \text{ overlap}\}$. There is a one-to-one correspondence between paths that visit each vertex of H at least once and DNA fragments that yield the given spectrum. The spectrum presented in Fig.1a yields a *path-graph* H_1. In this case, the immediate solution of the reconstruction problem is given by the fragment ATGCAGGTCC corresponding to the only path visiting all vertices of H_1. The spectrum shown in Fig.1b yields a more complicated graph H_2; however there still exists a unique solution ATGTGCCGCA to the reconstruction problem. For the spectrum presented in Fig.2 there are 2 *Hamiltonian* paths and 2 possible reconstructions.

For larger DNA fragments the overlap graphs become rather complicated. Bains and Smith [2], Lysov et al. [16] and Drmanac et al. [6] suggested several variants

of *backtracking* procedures for the reconstruction of a target DNA fragment from the spectrum. Unfortunately, these methods do not work for fragments that are hundreds of nucleotides long because of the high computational complexity of the Hamiltonian path problem.

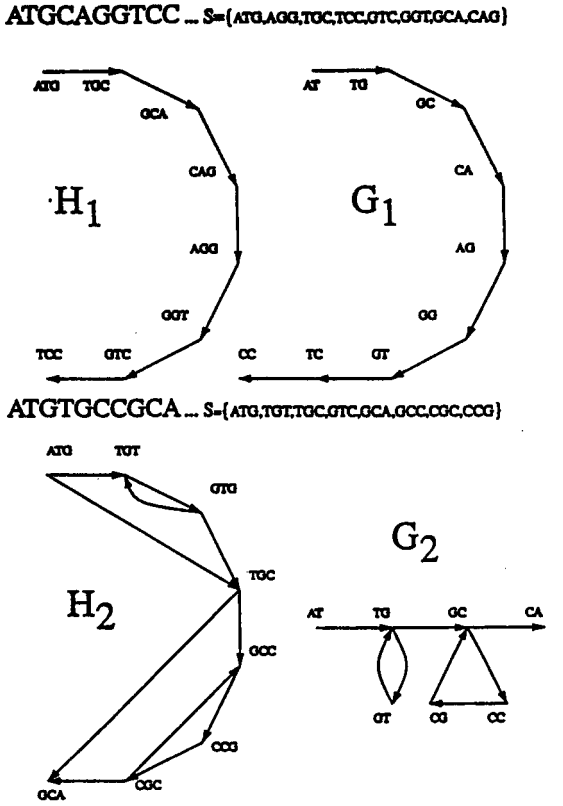

Figure 1. Examples of reductions of the sequences reconstruction problem to the Hamiltonian path problem (graphs H_1 and H_2) and to the Eulerian path problem (graphs G_1 and G_2).

2.2 SBH sequence reconstruction and Eulerian paths

Pevzner [20] reduced SBH reconstruction to the *Eulerian path* problem for which simple linear time algorithms are known. In this approach a graph G on the set of all $(l-1)$-tuples is constructed. An $(l-1)$-tuple v is joined by an arc with an $(l-1)$-tuple w if the spectrum contains an l-tuple for which first $l-1$ nucleotides coincide with v and the last $l-1$ nucleotides coincide with w (Fig.1 and 2). Each oligonucleotide from the spectrum corresponds to an *arc* in G but not to the *vertex* as in H. Therefore to find a target DNA fragment containing all oligonucleotides from the spectrum one has to find a path visiting all *arcs* of G, an *Eulerian path*. In contrast to the Hamiltonian path problem, the reduction to the Eulerian path in

the *de Bruijn* graph leads to simple linear time algorithms for SBH. In addition, the BEST theorem provides a formula for the number of possible reconstructions [20].

3 Biochemical, computer science and technological problems of SBH

The Eulerian path approach gives a complete treatment of the SBH reconstruction problem in the case of an *ideal* SBH experiments in which the *exact* count of the number of occurrences of each l-tuple in a target SBH fragment is known. However, even in this case, multiple Eulerian paths may exist, and we can not unambiguously reconstruct the DNA sequence from the spectrum.

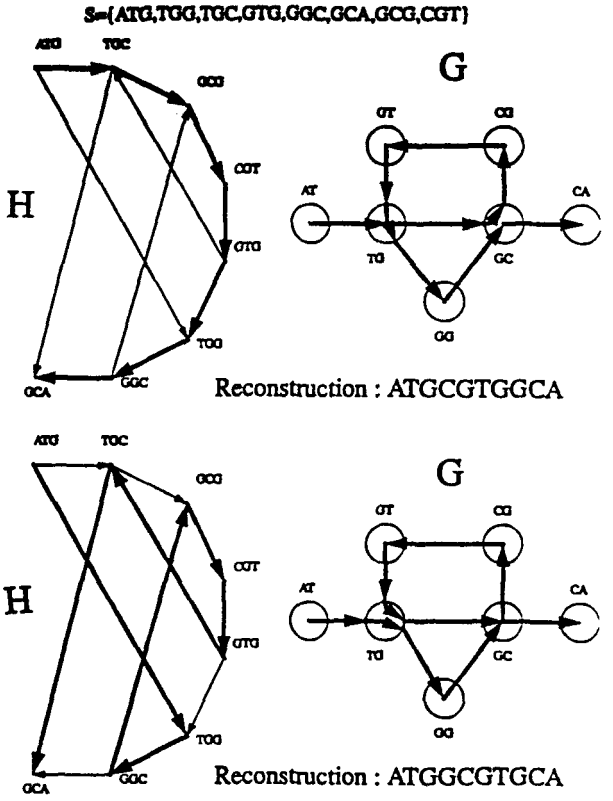

Figure 2. Spectrum S yields two possible reconstructions.

For *real* SBH experiments, the errors in the spectra that make reconstruction more complicated are unavoidable. Therefore, the problem of unambiguously reconstructing longer DNA fragments should be solved by the joint efforts of molecular biologists (reducing errors in the spectrum), computer scientists (reconstruction algorithms for spectra with errors and optimal chip design) and instrument designers

(increasing chip capacity and accurate detection of hybridization events). In the following sections we describe some problems concerning real SBH experiments and approaches to their solution.

3.1 Ambiguous Reconstruction and Additional Biochemical Experiments for SBH

Ambiguous reconstruction occurs when multiple DNA sequences have the same SBH spectrum (Fig.2). Pevzner et al., 1992 [22] demonstrate that even for a sequencing chip $C(10)$ containing all 4^{10} 10-tuples (such chip could be fabricated using the photolithographic technique of Affymetrix [10]) one can reliably decipher a DNA fragment only about 600 nucleotides long.

An example of ambiguity in sequence reconstruction is given in Fig.2. The graph G corresponding to the spectrum in Fig.2 contains a *branching vertex* TG. We don't know which 3-tuple (TGC or TGG) follows ATG in the original sequence. Therefore, we can not distinguish between correct and incorrect reconstructions If we could conduct an additional biochemical experiment (for example, hybridization of a target DNA fragment with 4-nucleotide ATGC) we would immediately find the correct reconstruction (the variant at the top of Fig.2 contains ATGC while the variant at the bottom of Fig.2 does not).

To analyze different additional biochemical experiments one needs a *characterization* of all DNA sequences with the given SBH spectrum. In the very first SBH studies the biologists described *string rearrangements* which do not change SBH spectrum and therefore do not allow one to unambiguously reconstruct these strings by SBH data (Drmanac et., 1989 [6]). However the problem how to describe *all* these rearrangements remained unsolved. Recently Ukkonen,1992 [25] conjectured that every two strings with the same SBH spectrum can be transformed into each other by the following transformations:

transposition If a string y can be written (in $(l-1)$-tuple notation) as

$$y = y_1 z_1 y_2 z_2 y_3 z_1 y_4 z_2 y_5$$

for some $(l-1)$ tuples z_1 and z_2 and for some strings $y_1, \ldots, y_5$ then the string $y = y_1 z_1 y_4 z_2 y_3 z_1 y_2 z_2 y_5$ where y_2 and y_4 have change places is called a *transposition* of y. If $y = y_1 z y_3 z y_4 z y_5$ where z is a $(l-1)$-tuple we also call $y = y_1 z y_4 z y_3 z y_5$ a transposition.

rotation If a string y can be written (in $(l-1)$-tuple notation) as $y = z_1 y_1 z_2 y_2 z_1$ for some $(l-1)$-tuples z_1 and z_2 and for some strings y_1 and y_2 then the string $y = z_2 y_2 z_1 y_1 z_2$ is called a *rotation* of y.

Trivially the above transformations do not change SBH spectrum. Pevzner, 1994 [23] demonstrated that every two strings with the same l-tuple composition can be transformed into each other by transpositions and rotations thus proving Ukkonen's conjecture.

The idea of using additional biochemical experiments to resolve branchings in the reconstruction process was proposed by Southern [24] (using a longer probe for each branching vertex) and Khrapko et al. [13] (continuous stacking hybridization).

Developing Southern's approach Gillevet [11] recently suggested using *genomic walking* [18] to resolve branchings (for the case of large-scale SBH experiments with rare and distant branching points).

Continuous stacking hybridization assumes an additional hybridization of short oligonucleotides which continuously extends duplexes formed by the target DNA fragment and the probes from the sequencing chip. In this approach additional hybridization with a short m-tuple on the chip $C(l)$ provides information about some $(l+m)$-tuples contained in the sequence. Computer simulations [13] suggest that continuous stacking hybridization with only 3 additional experiments provides an unambiguous reconstruction of a 1000 bp fragment in 97% of all cases. The questions regarding the computational complexity and the resolving power of these approaches need further studies.

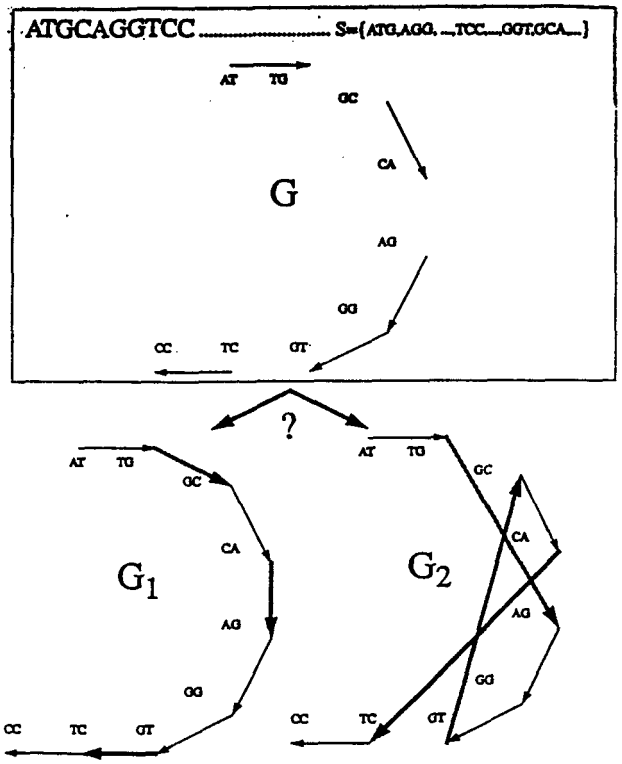

Figure 3. Two possible reconstructions for the case of incomplete hybridization (3 trinucleotide from the spectrum S are missing).

3.2 Incomplete hybridization

Because of the DNA secondary structure and other reasons, it is possible to lose information concerning some l-tuples in the course of hybridization (*false negative*). The problem is complicated by the fact that different oligonucleotides require different conditions for hybridization. In addition, the repeats of length l complicate

the analysis of hybridization intensity and lead to the incomplete spectra. If a DNA fragment has an l-tuple repeated, it will give a hybridization signal, but it is hard to determine the number of l-tuples present.

Thus the spectrum may not contain $n-l+1$ l-tuples, as in the case of ideal hybridization, but $n-l+1-k$ l-tuples, where k is the *defect* of the SBH experiment. Fig.3 presents the same spectrum as in Fig.1a with 3 trinucleotides missing. As a result the reconstruction process becomes more complicated (Fig.3 presents two possible reconstructions; only one of which is correct). However, Pevzner [21] presented an algorithm for SBH reconstruction from the spectrum with defects and demonstrated that for small defects the resolution is only slightly reduced compared with the ideal spectrum.

3.3 Non-specific hybridization

In the case of non-specific hybridization the spectrum contains l-tuples absent in a target DNA fragment (*false positive*). The solution to this problem requires biochemical and computer science methods. Lipshutz [15] proposed a *maximum likelihood* method for the SBH reconstruction problem and reduced SBH reconstruction to the *graph matching* problem. Given a spectrum and empirically derived rates of false positive and false negative hybridizations, he determines the most likely DNA fragment to have produced the spectrum. Anticipating that such a 'probabilistic reconstruction' does not seem satisfactory to biologists, note that classical DNA sequencing also yields a 'probabilistic reconstruction' [4].

Several biochemical approaches to the elimination of non-specific hybridization in SBH experiments have been proposed. These approaches allow one to build SBH detectors to better discriminate between perfectly matched and imperfectly matched oligonucleotides. Despite these recent advances in SBH biochemistry, hybridization data obtained by SBH are still much more ambiguous than computer scientists and biologists need them to be.

4 How to design sequencing chips?

Suppose that the number of position, m, on a sequencing chip is given and the problem is to devise m oligonucleotides (or m groups (*pools*) of oligonucleotides) to provide the maximum resolving power. In the very first SBH studies, Drmanac et al. [6] noticed that adding specific oligonucleotides to $C(8)$ significantly increased the ability to reconstruct targets. They devised a set of about 100,000 probes which contains $C(8)$ and also contains longer *self-overlapping* probes (like AAAAAAAAAAAA, ATATATATATAT or ATGATGATGATG). Bains and Smith [2] and Macevicz [17] have suggested *degenerative* probes (probes with positions that allow non-specific hybridization) to increase the chip's resolving power. No computational analysis of the merits and drawbacks of these approaches was presented.

Recently Bains [1] and Pevzner et al. [22] demonstrated that the classical $C(l)$ sequencing chips are redundant and therefore inefficient. They suggested a new family of chips which allows one to reduce the *capacity* (number of probes or pools of probes) of the chip by a factor of 5-15 times without significantly decreasing the

resolving power. Pevzner et al. [22] further raised the problem of devising *optimal* sequencing chips.

In order to discuss merits and demerits of these new chips we must redefine what we mean by a probe.

4.1 Generalized probes

We began by defining a probe to be a single oligonucleotide. Hybridization with that probe meant that the complement of the probe was a substring of the target. Let us expand the definition of a probe to be a *set* of oligonucleotides located at a single site on a chip. Now hybridization with a probe means that *at least one* oligonucleotide in the probe is complementary to a substring of the target. For example, WWS is a probe consisting of 8 trinucleotides

$$AAG, AAC, ATG, ATC, TAG, TAC, TTA, TTC$$

(W designates the *weak* nucleotides A or T, while S designates the *strong* nucleotides G or C). RYR is a probe consisting of 8 oligonucleotides

$$ATA, ATG, ACA, ACG, GTA, GTC, GCA, GCG$$

(R designates the *purines* A or G, while Y designates the *pyrimidines* T or C). $AAATXGGCA$ is a probe consisting of 4 oligonucleotides

$$AAATAGGCA, AAATTGGCA, AAATGGGCA, AAATCGGCA$$

(X designates any nucleotide A, T, G or C).

A *sequencing chip* is defined as a set of probes $C = (p_1, \ldots, p_{||C||})$. The *capacity* $||C||$ of the chip is the number of probes in C (every pool of oligonucleotides is counted as a single probe). Each DNA sequence F defines a subset of chip C consisting of the probes hybridizing with F (*spectrum* of F in C):

$$S(C, F) = \{p \epsilon C : \text{ probe } p \text{ contains an oligonucleotide occuring in sequence } \overline{F}\}$$

($\overline{F}$ stands for the sequence complementary to F). The spectrum of F in $C = (p_1, \ldots p_{||C||})$ is represented by the $||C||$-tuple vector $\mathbf{w}=(w_{(C,F)}(p_1), \ldots, w_{(C,F)}(p_{||C||}))$ where

$$w_{(C,F)}(p) = \begin{cases} 1, & \text{if an oligonucleotide from probe } p \text{ occurs in } \overline{F} \\ 0, & \text{otherwise} \end{cases}$$

A coordinate $w_{(C,F)}(p)$ of vector $\mathbf{w}$ indicates presence/absence of oligonucleotides from probe p in sequence $\overline{F}$.

4.2 Resolving power of chips

In order to compare different chips we need a measure of the resolving power. A fragment of length n (*n-fragment*) is called *unambiguously read by chip* C if there are no other n-fragments with the same spectrum in C. Let us consider an arbitrary chip C and define $D_n(C)$ to be the number of sequences of length n, which are ambiguously read by C. The probability of unambiguously deciphering a random sequence of the length n by means of the chip C (*resolving power* of chip C) is

$$p_n(C) = 1 - \frac{D_n(C)}{4^n}$$

While the resolving power is a rigorous measure of chip efficiency, there are neither theoretical results nor algorithms for estimating it. We introduce another definition of chip efficiency which does allow one to estimate the probability of unambiguous sequence reconstruction and to compare different chip designs.

Consider the sequence $F = X_1X_2 \dots X_{m-1}X_m, X_{m+1} \dots X_n$ and assume the first m nucleotides have already been determined. We will estimate the probability of unambiguously extending the sequence $F_m = X_1X_2 \dots X_m$ to the right by one nucleotide.

Since F_m is a possible reconstruction of the first m nucleotides of F then

$$S(F_m, C) \subset S(F, C).$$

There are four ways of extending F_m namely F_mA, F_mT, F_mG, F_mC. We define an extension of F_m by nucleotide N as a *possible extension* if

$$S(F_mN, C) \subset S(F, C) \tag{1}$$

Define $\epsilon(C, F, m)$ as

$$\epsilon(C, F, m) = \begin{cases} 0, & \text{if the condition (1) holds for exactly one of the four nucleotides} \\ 1, & \text{otherwise} \end{cases}$$

We call F *unambiguously extendable* after m with respect to chip C if $\epsilon(C, F, m) = 0$, otherwise F is called *ambiguously extendable* ($\epsilon(C, F, m) = 1$). The *branching probability* $q(C, n, m)$ is the probability of ambiguously extending a random n-sequence after the m-th nucleotide upon reconstruction with chip C. More precisely

$$q(C, n, m) = \frac{1}{4^n} \sum_F \epsilon(C, F, m)$$

where the sum is taken over all 4^n sequences F of length n.

Let us fix m and denote $q(C, n)$=$q(C, n, m)$. Obviously $q(C, n)$ is an increasing function of n. For a given *threshold branching probability* p, the maximum n satisfying the condition $q(C, n) \leq p$ is the maximal sequence length $n_{max}(C, p)$ allowing an *unambiguous reconstruction* with branching probability p. Below we demonstrate that for the chip $C(k)$, $n_{max}(C(k), p) \approx \frac{1}{3} \cdot \|C\| \cdot p$. For $k = 8$ and $p = 0.01$ it gives $n_{max} \approx 210$. In the following section we introduce new chips with $n_{max}(C, p) \approx \frac{1}{\sqrt{12}} \cdot \|C\| \cdot \sqrt{p}$. For $p = 0.01$ and the same capacity as the classical octanucleotide

chip $C(8)$, the new chips allow unambiguous reconstruction of the sequences of length $n_{max} \approx 1800$ nucleotides.

Comment. We emphasize that $n_{max}(C, p)$ (maximum fragment length for a given branching probability) and maximum fragment length for a given resolving power are very ***different*** characteristics of sequencing chips. The relations between these characteristics are still unclear. In particular, the claim that $n_{max}(C, 0.01) \approx 1800$ for the new sequencing chips does not mean that the resolving power of new sequencing chips $p_{1800}(C) \leq 0.01$.

4.3 Branching probability for classical chips $C(k)$

Consider the sequence $F = X_1X_2 \ldots X_{m-1}X_m, X_{m+1} \ldots X_n$ and assume that the first m nucleotides have been already determined. We estimate the probability of ambiguously extending the sequence $X_1X_2 \ldots X_m$ to the right by one nucleotide and compute these probabilities for $C(k)$ chips. Denote the last $(k-1)$-tuple in $X_1X_2 \ldots X_m$ as $V = X_{m-k+2} \ldots X_m$. Let Y_1, Y_2, Y_3 be three nucleotides different from X_{m+1}. For the sake of simplicity we suppose $m \geq k$ and $k \ll n \ll 4^k = ||C(k)||$

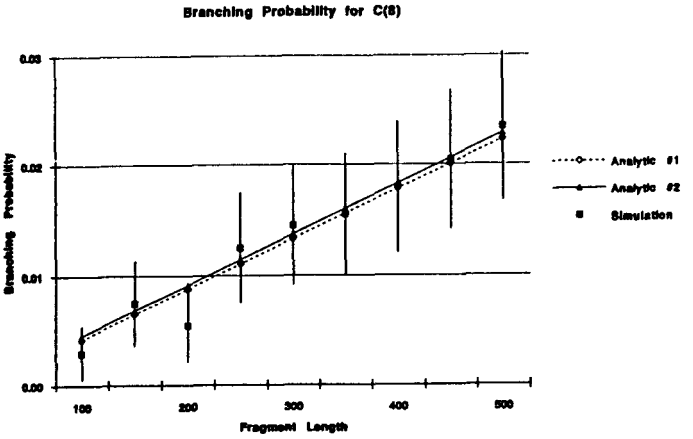

Figure 4. Simulations of sequence reconstructions with chip $C(8)$ based upon 2000 replications. Error bars are at $+/-2$ std dev. Analytic estimates #1 and #2 are given by $1 - ((1 - \frac{1}{4^k})^{n-k+1})^3$ and $\frac{3n}{\|C(k)\|}$ respectively.

The sequence F has an ambiguous extension after X_m using $C(k)$ if the spectrum $S(C_k, F)$ contains a VY_i k-tuples (here, Y_i is an arbitrary nucleotide different from X_{m+1}). In fact, for chip $C(k)$ the probability of ambiguous extension depends only on V: $q(C(k), n) = P\{VY_i \epsilon S(C_k, F)$ for i equals $1, 2$ or $3\}$. Assume that the probability of finding each of the 4^k k-tuples at a given position of F is equal to $\frac{1}{4^k}$. The probability that the spectrum of F does not contain VY_1 can be roughly estimated to be $(1 - \frac{1}{4^k})^{n-k+1}$ (we neglect the possibility of word self-overlapping

and marginal effects). The probability that the spectrum of F contains neither VY_1, nor VY_2 nor VY_3 can be estimated as $((1-\frac{1}{4^k})^{(n-k+1)})^3$. Therefore

$$q(C(k),n) = P\{VY_i \epsilon S(C_k,F)\} = 1 - P\{VY_1, VY_2, VY_3 \not\in S(C_k,F)\} \approx$$

$$1 - ((1-\frac{1}{4^k})^{n-k+1})^3 \approx \frac{3n}{4^k} = \frac{3n}{\|C(k)\|} \quad (2)$$

Assuming $q(C(k), n_{max}(C(k),p)) = p$ we derive

$$n_{max}(C(k),p) \approx \frac{1}{3} \cdot \|C(k)\| p$$

Note that n_{max} is linear in the chip capacity and p. (Fig.4).

5 New chips for SBH

We introduce 3 new chips for SBH and demonstrate their advantages over usual SBH chips $C(k)$.

5.1 Binary, gapped and alternating chips

The *binary chip* $C_{bin}(k)$ is the chip with capacity $\|C_{bin}(k))\| = 2 \cdot 2^k \cdot 4$ composed of all probes of two kinds

$$\underbrace{\{W,S\},\{W,S\},\ldots\{W,S\}}_{k},\{N\} \text{ and } \underbrace{\{R,Y\},\{R,Y\},\ldots\{R,Y\}}_{k},\{N\}$$

where R, S, W, Y are defined in section 4 and N is a specific base. Each probe is a mixture of 2^k oligonucleotides of length $k+1$. For example, the chip $C_{bin}(1)$ consists of the 16 probes

$$WA, WC, WG, WT, SA, SC, SG, ST, RA, RC, RG, RT, YA, YC, YG, YT,$$

each probe is a pool of two dinucleotides.

The *gapped chip* $C_{gap}(k)$ is the chip with capacity $\|C_{gap}(k))\| = 2 \cdot 4^k$ composed of all probes of two kinds

$$N_1N_2\ldots N_k \text{ and } N_1N_2\ldots N_{k-1}\underbrace{XX\ldots X}_{k-1}N_k$$

where N_i is a specific base and X is an unspecified base as above. Each probe of the first kind consists of the only oligonucleotide of length k, each probe of the second kind consists of 4^{k-1} oligonucleotides of length $2k-1$. Chips similar to the gapped chip were proposed in [14] and [8].

The *alternating chip* $C_{alt}(k)$ is the chip with capacity $\|C_{alt}(k))\| = 2 \cdot 4^k$ composed of all probes of two kinds

$$N_1XN_2X\ldots N_{k-2}XN_{k-1}XN_k \text{ and } N_1XN_2X\ldots N_{k-2}XN_{k-1}N_k.$$

Each probe of the first kind consists of the 4^{k-1} oligonucleotide of length $2k-1$, while each probe of the second kind consists of 4^{k-2} oligonucleotides of length $2k-2$.

5.2 Branching probabilities of new chips

Now we estimate $q(C_{bin}(k), n)$ for $m \geq k+1$ and $n \ll ||C_{bin}(k)||$. In this case the ambiguity arises when the spectrum $S(C_{bin}(k), F)$ contains *both* a $V'Y_i$ probe and a $V''Y_i$ probe (here $Y_i \neq X_{m+1}$, V' is V written in the $\{W, S\}$ alphabet, V'' is V written in the $\{R, Y\}$ alphabet). Assume that the probability of finding each $k+1$-tuple in $C_{bin}(k)$ at a given position of F is equal to $\frac{1}{4 \cdot 2^k}$ and neglect self-overlaps. Then the probability that the spectrum of F does not contain $V'Y_1$ can be roughly estimated as $(1-\frac{1}{4 \cdot 2^k})^{n-k}$. Therefore the probability that the spectrum of F contains both $V'Y_1$ and $V''Y_1$ is

$$(1-(1-\frac{1}{4 \cdot 2^k})^{n-k}) \cdot (1-(1-\frac{1}{4 \cdot 2^k})^{n-k} \approx \frac{n^2}{4 \cdot 2^k \cdot 4 \cdot 2^k}.$$

Similarly to (2) we derive:

$$q(C_{bin}(k), n) = P\{V'Y_i \epsilon S(C_{bin}(k), F) \text{ and } V''Y_i \epsilon S(C_{bin}(k), F)\} \approx$$

$$1-(1-\frac{n^2}{4 \cdot 2^k \cdot 4 \cdot 2^k})^3 \approx 3 \cdot \frac{n}{4 \cdot 2^k} \cdot \frac{n}{4 \cdot 2^k} = \frac{12n^2}{||C_{bin}(k)||^2}$$

Therefore for $C_{bin}(k)$

$$n_{max}(C_{bin}(k), p) \approx \frac{1}{\sqrt{12}} \cdot ||C_{bin}(k)|| \sqrt{p}.$$

Note that n_{max} is still linear in C_{bin} but, now grows as the square root of p (Fig.5).

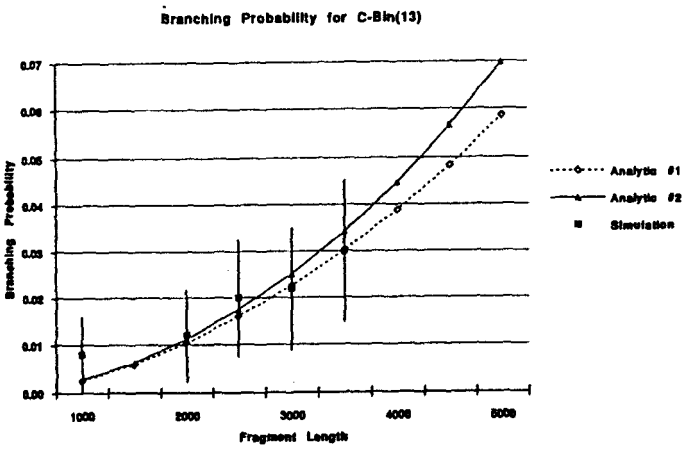

Figure 5. Simulations of sequence reconstructions with binary chip $C_{bin}(13)$ (the same capacity as $C(8)$) based upon 2000 replications. Error bars are at $+/-2$ std dev. Analytic estimates #1 and #2 are given by $1-(1-(\ (1-(1-\frac{1}{4 \cdot 2^k})^{n-k}) \cdot (1-(1-\frac{1}{4 \cdot 2^k})^{n-k}\ )^3$ and $\frac{12n^2}{||C_{bin}(k)||^2}$ respectively.

Next we estimate the branching probability of gapped chips $C_{gap}(k)$

Let $m \geq 2k-1$ and $n \ll ||C_{gap}(k)||$. Denote $U = X_{m-2k+4} \ldots X_{m-k+2}$. In this case the ambiguity arises when the spectrum $S(C_{gap}(k), F)$ contains *both* a VY_i k-probe and a $U\underbrace{XX\ldots X}_{k-1}Y_i$ $(2k-1)$-probe (here $Y_i \neq X_{m+1}$). Assume that the probability of finding each $k+1$-tuple in $C_{gap}(k)$ at a given position of F equals $\frac{1}{4^k}$ and neglect self-overlaps. Then the probability that the spectrum of F does not contain VY_1 can be roughly estimated as $(1-\frac{1}{4^k})^{n-k}$. The probability that the spectrum of F does not contain $U\underbrace{XX\ldots X}_{k-1}Y_1$ can be roughly estimated as $(1-\frac{1}{4^k})^{n-(2k-1)+1}$. Therefore the probability that the spectrum of F contains both VY_1 and $U\underbrace{XX\ldots X}_{k-1}Y_1$ is

$$(1-(1-\frac{1}{4\cdot 2^k})^{n-k})\cdot(1-(1-\frac{1}{4\cdot 2^k})^{n-2k+2}) \approx \frac{n^2}{4\cdot 2^k\cdot 4\cdot 2^k}$$

Similarly to (2) we derive:

$$q(C_{gap}(k),n) = P\{VY_i \epsilon S(C_{gap}(k),F) \text{ and } UY_i \epsilon S(C_{gap}(k),F)\} \approx$$

$$1-(1-\frac{n^2}{4\cdot 2^k\cdot 4\cdot 2^k})^3 \approx 3\cdot\frac{n}{4\cdot 2^k}\cdot\frac{n}{4\cdot 2^k} = \frac{12n^2}{|C_{gap}(k)|^2}$$

Therefore for $C_{gap}(k)$

$$n_{max}(C_{gap}(k),p) \approx \frac{1}{\sqrt{12}}\cdot||C_{gap}(k)||\sqrt{p}$$

Similar arguments demonstrate that

$$n_{max}(C_{alt}(k),p) \approx \frac{1}{\sqrt{12}}\cdot||C_{alt}(k)||\sqrt{p}$$

Reconstruction of a DNA fragment from its spectrum over the new chips is a difficult problem. Let us confine ourselves to the case of binary chips and assume that the first k nucleotides of the DNA fragment are known. This allows us to start reconstruction process; otherwise we should try all possible k-tuples (compare with [1]). We will be extending the DNA fragment from the starting k-tuple by a nucleotide N checking the condition (1). In cases where the current DNA fragment is ambiguously extendable we apply a backtracking procedure. A more efficient solution to this problem is not yet known.

6 Conclusion

SBH has certain advantages over the current DNA sequencing technology, in particular

- SBH has the potential to sequence long DNA fragments in a single experiment
- SBH is amenable to automation [10, 7, 9]

– Small-scale SBH can be used for physical mapping and clinical diagnostics [3].

In our opinion the primary obstacle in SBH implementation is high fidelity determination of hybridization. Recent breakthroughs in SBH have gone a long way towards demonstrating the technical feasibility of generating the data necessary to perform sequencing by hybridization. Parallel combinatorial algorithms allow efficient DNA reconstruction from hybridization data even in the presence of possible hybridization errors. These advances indicate that SBH is close to large scale implementation although it is not yet at a stage where a single approach can be judged most promising. Remaining problems will require continued close collaboration among biochemists, computer scientists, and instrument designers. In the area of chip design, new concepts may allow the use of chips with much fewer probes that are still able to unambiguously reconstruct long DNA fragments.

7 Acknowledgements

We wish to thank Patrick Gillevet and Michael Waterman for helpful discussions.

References

1. Bains W. Hybridization methods for DNA sequencing. *Genomics*, **11**, (1991), 294-301
2. Bains W., Smith G.C. A novel method for DNA sequence determination. *J.Theor. Biol.*,**135**, (1988), 303-307
3. Cantor C., Mirzabekov A., Southern E. SBH: An idea whose time has come. *Genomics*.**13** (1992), 1378-1383
4. Churchill G.A., Waterman M.S. The accuracy of DNA sequences: estimating sequence quality. *Genomics*. **14** (1992), 89-98.
5. Drmanac R., Crkvenjakov R. *Yugoslav Patent Application 570.* (1987)
6. Drmanac R.,Labat L.,Brukner I.,Crkvenjakov R. Sequencing of megabase plus DNA by Hybridization. *Genomics*,**4** (1989) ,114-128
7. Drmanac S., Labat I., Crkvenjakov R., Vicentic A., Gemmell A., Drmanac R. Sequencing by hybridization (SBH): a production line to sequence one million M13 clones arrayed on membranes. *Electrophoresis*, **13**, (1992), 566-573
8. Drmanac R., Crkvenjakov R. m Sequencing by hybridization (SBH), with oligonucleotides probes as an integral approach for the analysis of complex genomes. *International Journal of Genome Research*, **1**, (1992), 59-79
9. Eggers M., Beattie K., Shumaker J., Hogan M., Hollis M., Murphy A., Rathman D., Erlich D. Genosensors: microfabricated devices for automated high throughput DNA sequence analysis. In 'Genome Mapping and Sequencing', (Abstracts of the paper presented at the 1992 meeting arranged by M.Olson, C.Cantor and R.Roberts), Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, (1992), 111
10. Fodor S.P.A., Read J.L., Pirrung M.S., Stryer L., Lu A.T., Solas D. Light-directed spatially addressable parallel chemical synthesis. *Science*, **251**, (1991) 767-773
11. Fodor S.P.A., Rava R.P., Huang X.C., Pease A.C., Holmes C.P., Adams C.L. Multiplex biochemical assays with biological chips. *Nature*, **364**, (1993) 555-556
12. Gillevet P.M. Mutliplex genomic walking: Integration of the wet lab and computer lab into a single prototyping environment. In C.Cantor, J.Fickett. H.Lim, R.Robbins (eds.) *The Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, (1992), 197-206

13. Khrapko K.R.,Lysov Yu.P.,Khorlin A.A.,Shik V.V.,Florent'ev V.L.,Mirzabekov A.D. An oligonucleotide apporach to DNA sequencing. *FEBS Letters*, **256**, (1989), 118-122
14. Khrapko K.R.,Lysov Yu.P.,Khorlin A.A., Ivanov I.B., Yershov G.M., Vasilenko S.K., V.V.,Florent'ev V.L.,Mirzabekov A.D. A method for DNA sequencing by hybridization with oligonucleotide matrix. *DNA Sequence*,**1**, (1991), 375-388
15. Lipshutz R.J. Maximum likelihood DNA sequencing by hybridization. *J. Biom. Struct. Dyn.* **11**, (1993), 637-653
16. Lysov Yu.P., Florent'ev V.L.,Khorlin A.A.,Khrapko.,Shik V.V., Mirzabekov A.D. DNA Sequencing by hybridization with oligonucleotides. A novel method. *Dokl. Acad. Sci USSR*, **303**, (1988) 1508-1511
17. Macevicz S.C. *International Patent Application PS US89 04741* (1989)
18. Maskos U., Southern E.M. *Nucleic Acids Research*, **20**, (1992), 1675-1681
19. Ohara, O. Dorit, R., Gilbert W. Direct genomic sequencing of Bacterial DNA: The pyruvate kinase I gene of Eschericia coli. *Proc Natl. Acad. Sci. USA*, **86**, (1989), 6883
20. Pease A.C., Solas D., Sullivan E.J., Cronin M.T., Holmes C.P., Fodor S.P.A. Oligonucleotide arrays for for rapid DNA sequence analysis. *Proc. Natl. Acad. of Sci. USA*, **91**, (1994) 5022-5026
21. Pevzner P.A. l-tuple DNA sequencing: a computer analysis. *J.Biom.Struct. and Dyn.*, **7**, (1989) 63-73
22. Pevzner P.A., Lysov Yu.P., Khrapko K.R., Belyavsky A.V., Florentiev V.L., Mirzabekov A.D. Optimal chips for megabase DNA sequencing. *J.Biomol.Struct.Dyn.*, **9**, (1991) 399-410
23. Pevzner P.A. DNA physical mapping and alternating Eulerian cycles in coloured graphs. *Algorithmica*, **12**, (1994) (to appear)
24. Southern E. *United Kingdom Patent Application GB8810400.* (1988)
25. Ukkonen E. Approximate string matching with q-grams and maximal matches. *Theoretical Computer Science*, **92**, (1992) 191-211