Optimizing the BAC-End Strategy for Sequencing the Human Genome

Richard M. Karp* Ron Shamir†

April 24, 1999

Abstract

The rapid increase in human genome sequencing effort and the emergence of several alternative strategies for large-scale sequencing raise the need for a thorough comparison of such strategies. This paper provides a mathematical analysis of the BAC-end strategy of (Venter,1996) showing how to obtain an optimal choice of parameters. The analysis makes very mild assumptions. In particular, it accommodates variable clone length and inhomogeneity of the distribution of clone locations. The analysis implies that the BAC-end strategy is very close to optimal in terms of cost, under a wide range of experimental scenarios.

*Department of Computer Science and Engineering, University of Washington, Box 352350, Seattle, Washington 98195-2350.

†Department of Computer Science, School of Mathematics, Tel Aviv University, Tel Aviv, 69978, Israel.

1 Introduction

With the Human Genome Project moving from the mapping phase into the sequencing phase, the challenge of improving the efficiency of large scale sequencing has become central. The classical strategy set forth by the founders of the Human Genome Project (Office of Technology Assessment, 1988), (National Research Council, 1988) has been to first construct a clone map and then extract from it a set of clones for sequencing, covering the genome with minimal overlap. Two recent proposals suggest alternative strategies that bypass the mapping stage. These strategies involve fewer laboratory procedures and are therefore more automatable. The first of these is the BAC-end strategy proposed in 1996 by Venter, Smith and Hood (Venter, 1996). Quite recently, Venter et al. (Venter, 1998) proposed to perform a complete direct shotgun sequencing of the whole genome, using several different clone types, with the library of end-sequenced BAC clones serving as a scaffold for the process, as in the BAC-end strategy. A thorough comparison of these strategies is needed. Mathematical analysis is an essential component of this comparison along with simulation studies and pilot experimental projects. Here we provide a mathematical analysis of the BAC-end strategy, showing how to obtain an optimal choice of parameters. Our analysis makes very mild assumptions. In particular, it accommodates variable clone length and inhomogeneity of the distribution of clone positions.

The protocol of the method is described below. For concreteness, the values of the relevant parameters are given as in (Venter, 1996), although they may vary and will be changed later.

1. A library of BAC clones with relatively high redundancy is generated. The average BAC clone has length of 150,000 bases, so 300,000 human genome clones obtain redundancy of about 15.

2. Both ends of each BAC are sequenced, obtaining a database of 600,000 sequences. These sequences (of typical length 500 bases) are scattered on average every 5,000 bases along the genome. These sequences are called *sequence tagged connectors*, or STCs.

3. Each BAC clone is fingerprinted, using restriction enzymes.

4. A seed BAC for each region of interest (e.g., a chromosome) is chosen and fully sequenced. This can be done, for example, using the conventional shotgun sequencing strategy with M13 or plasmid clones.

5. The region already sequenced is extended by sequencing clones that overhang it to the right and left. We describe the extension to the right, with the extension to the left being similar. Let R be the rightmost clone in the region already sequenced. By comparing the sequence of R to the database of STCs, on average 30 STCs are identified that match subsequences within R.

Their BAC clones overlap R. A clone showing minimal overlap with R, and demonstrating internal consistency via comparison of its fingerprints to other clones, is chosen and sequenced next.

The last step is repeated until the whole target is sequenced.

Both the BAC-end strategy described here and the classical strategy based on physical mapping seek to select for sequencing a set of clones that cover the genome with minimal overlap. In principle the BAC-end strategy requires fewer and simpler laboratory procedures and avoids the need to create a complete physical map before sequencing can proceed. See (Venter, 1996) for a discussion of many other advantages.

2 Discussion and Results

Recently, A. Siegel and colleagues (Siegel, 1998) have performed a statistical analysis of the cost of sequencing the whole genome by the STC approach. Here we study a more general model that accommodates variable clone length and inhomogeneity of the distribution of clone locations. Within this model we determine closed-form expressions for the parameter values that minimize the expected cost of the entire sequencing process.

Let us first fix some terminology. We concentrate on a single target region of interest, such as a chromosome. In the midst of the process, the set of BAC clones sequenced so far constitutes a contiguous segment (contig) of the target. Without loss of generality we consider a step in which the sequenced contig is about to be extended to the right. We call the most recently sequenced (rightmost) BAC the *frontier*. Those BACs that have their right endpoint to the right of the contig and their left endpoint in the contig are called *overhanging*. We call a clone *bad* if it is artifactual, *i.e.*, if it has a sequence that differs from any contiguous segment in the true target sequence, due to rearrangements, deletions, chimerism, etc.

We make several assumptions on the progress of the sequencing process: Our first assumption is that the process will not "get stuck". More precisely, after each clone sequencing step, there is an overhanging clone that can be used for the next step. Note that, except for end effects, this assumption is equivalent to assuming that the collection of clones covers the whole target without gaps. This is a reasonable working assumption if the redundancy of the clones is sufficiently high, as then gaps will be infrequent.

The second assumption is that the choice of the next frontier clone is always correct, in the sense that a bad clone is never chosen, nor is a good clone that does not overhang the current frontier. This holds with high probability if the fingerprint screening is stringent enough, and if a very high level of similarity between the clone's STC and the sequence of the current frontier is required. The analysis of Siegel et al., which takes into consideration the highly repetitive nature of human DNA sequences, indicates that the effort wasted on sequencing bad or misplaced clones is negligible.

2.1 An Optimal Strategy

Our analysis of the BAC-end strategy will be in terms of the cost involved. The key parameter to be chosen is the number of BAC clones. The optimal choice of this parameter involves a trade-off between the cost of BAC preparation and the subsequent cost of fully sequencing a subset of the BACs. In this section we give the broad outlines and main conclusions of the analysis. Mathematical details are given in Section 3.

We shall denote the average preparation cost of steps 1-3 for a BAC in the library by α. This cost includes library construction, sequencing both ends of the BAC to generate its STCs, fingerprinting, computation and material handling. The average cost of fully sequencing a BAC (step 4 or 5) is denoted by β. The sequencing of each STC is done by a single reaction and thus the STC sequencing cost per base is much (5-10 times) cheaper than that of of the finished sequence, where high accuracy must be achieved by resequencing each base several times in different subclones. A much less accurate STC sequence suffices for making the right connection in step 5. An important parameter is the ratio β/α.

2.1.1 Variable clone length

Our model for clone distribution is as follows: The target is a contiguous stretch of length N, denoted by the interval

$[0, N]$. The left endpoints of clones are uniformly distributed in the interval $[0, N]$. (Since N is sufficiently large compared to any single clone's length, the impact of clones overhanging the right end of the target is negligible.) Clone lengths are bounded and have a distribution with cumulative distribution function F and expectation μ. The *redundancy* (sum of clone lengths divided by N) is R. Under these assumptions we may take the distribution of left end points of clones to be Poisson with rate R. Such a Poisson model has been demonstrated to match quite closely with experimental observations (Lander, 1988). We allow the cost of sequencing to change non-linearly with clone length (typically the cost increases more than linearly with length). The cost of sequencing a clone of length x is denoted by $C(x)$, so that the expected cost of sequencing a clone is: $\beta = E[C] = \int C(x)dF(x)$.

Denote by Y_i the length of the newly sequenced segment of the i-th frontier, i.e., the progress made in one repetition of step 5. Y_i is a random variable whose expectation is

$$E[Y] = \mu(1 - \frac{1}{R}),$$

as $\frac{\mu}{R}$ is the expected distance from the right end of the current frontier to the first left end on its left. Let $\nu(t)$ be the number of clones needed to sequence the whole target of length t. In other words, $\nu(t)$ is the least integer k such that $\sum_{i=1}^{k} Y_i \geq t$. Using results from Section 3 about Renewal-Reward processes we find that the expected total sequencing cost is, up to a small additive constant,

$$E[\sum_{i=1}^{\nu(N)} C_i] = \frac{NE[C]}{E[Y]} \quad (1)$$

Hence, the expected total cost of the project is (up to a small additive constant) $\alpha\frac{RN}{\mu} + \beta\frac{N}{\mu(1-\frac{1}{R})}$. Renormalizing N so that $\mu = 1$, this value is

$$N(\alpha R + \beta\frac{R}{R-1}).$$

The optimal redundancy R_{opt} is thus obtained when $\alpha - \beta(R_{opt}-1)^{-2} = 0$, or

$$R_{opt} = \sqrt{\frac{\beta}{\alpha}} + 1$$

and the total optimal cost is

$$N(\sqrt{\alpha\beta} + \alpha + \beta\frac{\sqrt{\frac{\beta}{\alpha}}+1}{\sqrt{\frac{\beta}{\alpha}}}) = N(\sqrt{\beta}+\sqrt{\alpha})^2.$$

Consider a hypothetical *ideal project* in which no end sequencing is required and the clones selected for sequencing cover the whole target without overlap. The expected cost of the ideal project is βN. This quantity is a lower bound on the expected cost of any actual project. Hence, the cost of the optimal BAC-end strategy is larger than what is ideally possible by a factor of at most $\frac{(\sqrt{\beta}+\sqrt{\alpha})^2}{\beta} = (1+\sqrt{\frac{\alpha}{\beta}})^2 = \frac{\alpha}{\beta}R_{opt}^2$.

Figure 1 shows the expected cost of a sequencing project using the BAC-end strategy relative to the cost of an ideal project. Each curve corresponds to a different value of α/β. Note that the optimal redundancy decreases with α/β. Note also that the impact of changing α/β in the range $1/1000 \leq \alpha/\beta \leq 1/2000$ is very modest. For α/β in that range, the impact of using a suboptimal redundancy (say, within a range

± 10 from the optimal redundancy) is very minor, as all the corresponding cost curves are very flat near their optimum redundancy.

To put the result in real dollar terms, the values in the analysis of Siegel et al. were used, namely, $N = 20,000$, $\alpha = \$48$, $\beta = \$67,500$. The optimal redundancy is then 38.5, and the overall cost is \$1.423 billion. The cost of "an ideal project" with these parameters would be \$1.350 billion. Hence, the BAC-end strategy is within five percent of what may be achieved by any conceivable sequencing strategy. Figure 1 shows that this upper bound on the percentage of waste compared to an ideal project is quite insensitive to changes, in the realistic range where $1/1000 \leq \alpha/\beta \leq 1/2000$. These results are in agreement with the results given by Siegel et al. for the case where all clones are of the same length. Both the sequencing cost and the cost of an ideal project may be substantially reduced in view of the recently announced progress in sequencing technology (Venter, 1998), but the key quantity $\frac{\alpha}{\beta}$ will remain in the same range, and thus our conclusions about the near-optimal efficiency of the BAC-end strategy will not change.

2.1.2 Variable clone density

We now consider the situation where the distribution of clones is not uniform across the target DNA. We assume that the left end points of the clones are drawn independently from a common probability distribution with density $g(x)$ over the interval $[0, N]$ (as before, we ignore the impact of clones overhanging the right end of the target). We assume that the distribution of the length of a clone is independent of its left end point, and that the length scale is normalized so that μ, the expected length of a clone, is 1.

We assume that the target DNA is divided into a finite number of intervals, such that the probability density $g(x)$ within each interval is constant. We denote the number of intervals by m, the length of the ith interval by N_i, and the probability density within the ith interval by p_i; thus $\sum_{i=1}^{m} N_i p_i = 1$. We assume that m is small compared to N, the length of the target.

Let the number of clones be NR, where R is the average redundancy across the entire target. The expected redundancy R_i in the ith interval is NRp_i. Applying the Renewal-Reward Theorem (cf. Section 3) to each interval in the same manner as it was applied to the entire target in the previous subsection, we find that, up to a negligible error proportional to m, the expected cost of the project is $\sum_{i=1}^{m} N_i(\alpha R_i + \beta \frac{R_i}{R_i - 1})$. The optimal redundancy R_{opt} is obtained when $\sum_{i=1}^{m} N_i p_i(\alpha - \beta(NR_{opt}p_i - 1)^{-2}) = 0$. The value of R_{opt}, and hence the minimum expected total cost of the project, can be determined numerically from this relation. In the parameter range of practical interest R_i will be large enough in each interval that the effect of gaps in the clone coverage is negligible. Under this assumption the term $\beta \frac{R_i}{R_i - 1}$ in the expected cost is closely approximated by

$\beta(1+\frac{1}{R_i})$, leading to the following approximate expression for the expected total cost:

$$\sum_{i=1}^{m} N_i(\alpha R_i + \beta + \frac{\beta}{R_i}).$$

This expression is minimized at the point

$$R_{approx} = \frac{1}{N}\sqrt{\frac{\beta}{\alpha}\sum_{i=1}^{m}\frac{N_i}{P_i}}.$$

The quantity R_{approx} is a close approximation to the optimal redundancy.

To illustrate the effect of a nonuniform clone distribution, we considered the case where the target is divided ito two intervals of equal length, with uniform clone density in each interval. Figure 2 demonstrates that the effect of inhomogeneity of clone density on the optimal cost is quite small unless the ratio of probability densities between the two intervals is very large or the ratio $\frac{\alpha}{\beta}$ is much larger than one would expect in practice.

3 Methods

We require some facts from renewal theory. Let $\{Y_i\}$ and $\{C_i\}$ be sequences of bounded, positive random variables such that the pairs (Y_i, C_i) are mutually independent and identically distributed but, for any given i, Y_i and C_i need not be independent. Let $\nu(t)$ be the least n such that $\sum_{i=1}^{n} Y_i > t$. Let $C(t) = \sum_{i=1}^{\nu(t)} C_i$. Wald's equation (Grimmet, 1992) yields the following result, known as the Renewal-Reward Theorem:

$$E[C(t)] = E[\sum_{i=1}^{\nu(t)} Y_i]\frac{E[C_1]}{E[Y_1]}.$$

In order to derive equation (1) from these results we need the following *high-redundancy approximation*: For any i, let F_i denote the ith frontier clone encountered as the sequencing process progresses to the right. Let a_i and b_i respectively denote the left and right end points of F_i. Then $a_{i+1} \leq b_i$, and the quantity $b_i - a_{i+1}$ is exponentially distributed with rate R, irrespective of the past history of the process. Because the left end points of clones are distributed according to a Poisson process of rate R, this approximation would be exact if the length of the interval $[b_{i-1}, b_i]$ were infinite. Since the interval is finite the possibility exists that no clone has its left end in the interval $[b_{i-1}, b_i]$, in which case our approximation deviates from reality. This anomalous situation is very rare when the redundancy R is high, since the length of this interval is expected to be very large compared to $1/R$, the expected distance that has to be traversed leftward from the right end of F_i before encountering the left end of a clone.

Recall that Y_i denotes the progress made in sequencing the ith frontier clone as the BAC-end sequencing process advances to the right and C_i denotes the cost of sequencing the ith frontier clone. Under the high redundancy approximation the pairs (Y_i, C_i) satisfy the conditions of the Renewal-Reward Theorem, and $C(t)$ denotes the total cost of sequencing an interval of length t (with the convention that the entire cost of sequencing the first clone that overhangs the interval to the right is included in $C(t)$). From the Renewal-Reward Theorem, since the Y_i and C_i are bounded above and below, it follows that $E[C(t)]$ differs from $\frac{tE[C_1]}{E[Y_1]}$ by a quantity that is uniformly bounded, independent of t. This justifies Equation 1.

4 References

Grimmet, G. R. and D.R. Stirzaker. 1992. Probability and random processes. Oxford University Press.

Lander, E. S. and M.S. Waterman.1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231.

National Research Council Report.1988. Mapping and sequencing the human genome.National Academy Press.

Office of Technology Assessment, U. S. Congress.1988. Mapping our genes – the genome projects, how big, how fast?. *Technical Report OTA-BA-373.*

Siegel, A. F., Trask, B., Roach, J., Mahairas, G. G., Hood, L., and G. van den Engh. 1998.Analysis of sequence-tagged-connector strategies for DNA sequencing.

Venter, J. C., Smith, H. O., and L. Hood. 1996. A new strategy for genome sequencing. *Nature* **381:** 364-366.

Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O. , and M. Hunkapiller.1998. Shotgun sequencing of the human genome. *Science.* **280:** 1540-1542.

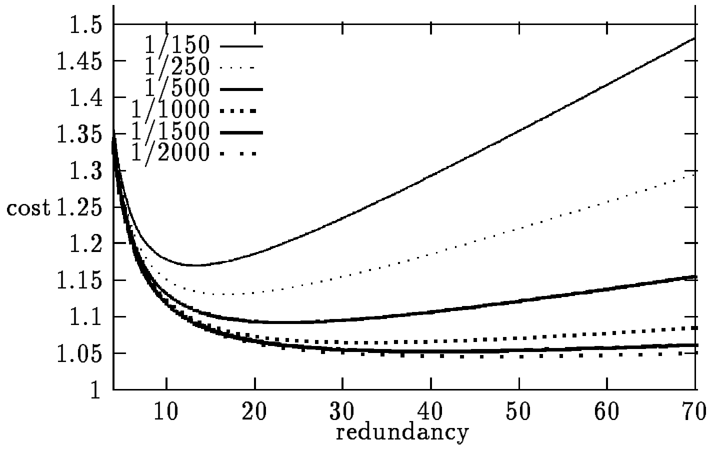

Figure 1: Overall project cost, relative to the "ideal project" cost. The different plots correspond to different values of α/β. The keys refer to the plots from top to bottom.

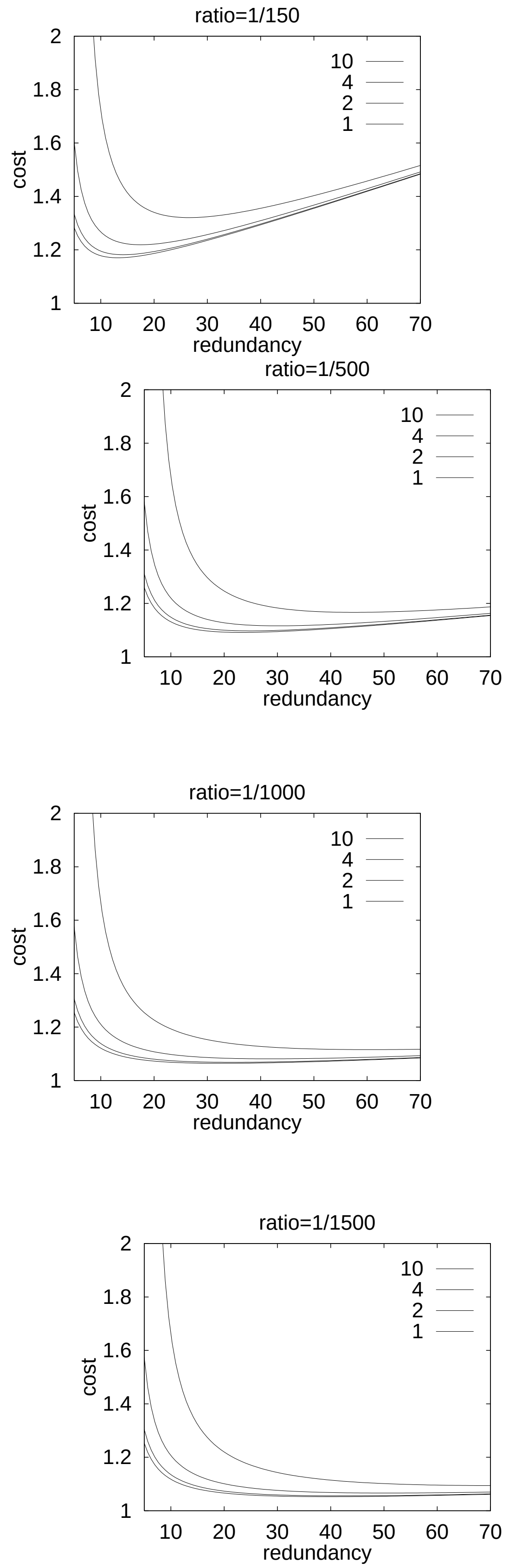

Figure 2: Impact of the inhomogeneity on the total cost. Y axis: cost relative to the "ideal cost". Model used: Two halves of the genome with different redundancies. Each figure refers to a different α/β ratio. The lines in each figure plot the cost for different redundancy ratios. The keys refer to the plots from top to bottom.