

# Sequencing by Hybridization – A Simulation Study of Performance on Genomic Sequences

Doron Lipson<sup>1</sup>, Ziv Nevo, Ari Frank, Dolev Dotan, Zohar Yakhini<sup>2</sup>  
Computer Science Department, Technion, Haifa, Israel

**Keywords:** Sequencing, Hybridization, Micro-arrays, Markov models

## 1 Introduction.

Sequencing by Hybridization (SBH)[1,2] is a theoretical method for de-novo sequencing of DNA by means of reconstruction of the sequence from its hybridization pattern. Typically, arraying a complete set of  $k$ -mers is considered, although different setups, such as degenerate probe arrays, are also possible [3,5]. While this method is not competitive with current biochemical sequencing methods, variations of the model may be practical for problems such as SNP genotyping.

Formally, given a DNA sequence  $S$  of length  $n$  and a probe length  $k$  (where  $k \ll n$ ), hybridization of  $S$  on an array of all possible  $k$ -mers defines a hybridization pattern  $\sigma_k(S)$ . Given  $\sigma_k(S)$ , the length  $n$ , and the  $k$ -prefix and  $k$ -suffix of  $S$ , SBH algorithms attempt to reconstruct the sequence  $S$  [4].

Even when assuming perfect and stringent hybridization SBH still suffers from an inherent information problem. For any given  $k$ , there exist sequences that cannot be uniquely reconstructed by SBH. Specifically, sequences that contain a double alternating repeat of  $(k-1)$ -mers ( $abc\beta cad\beta e$ ) or a triple repeat ( $abc\gamma c\gamma d$ ) cannot be uniquely reconstructed. Therefore, for a given  $k$ , the probability of success of SBH in uniquely reconstructing randomly generated sequences decreases rapidly as sequence length,  $n$ , increases. Since, from the information point of view,  $(k-1)$ -mer repeats are the basis of the failure mechanism we expect performance to be even worse on genomic sequences.

In this report, we describe an analysis by simulations of SBH performance on genomic sequences. We specifically test the extent to which increasing order Markov models of the data explain SBH poor performance on genomic sequences.

## 2 Method.

The SBH algorithm was implemented for 3 different values of probe length  $k = 8, 10, 12$ . For each  $k$ , sequences of length  $n$  in the range 100-10000 were tested. The success rate of unique reconstruction of the original sequence was determined for each combination of  $(k, n)$  over 1000 different DNA sequences. These sequences were randomly generated or randomly selected from the genomic databases of: a) *S. cerevisiae*, b) *E. coli* and c) *H. influenza*.

## 3 Results.

To facilitate comparison, we follow earlier work on analyzing SBH performance [4] and define  $n_{90}(k)$  to be the length  $n$  at which the success rate drops below 0.9. For randomly generated data we measured:  $n_{90}(8) = 200$ ,  $n_{90}(10) = 800$ ,  $n_{90}(12) = 3100$ , an exponential dependence on  $k$ .

---

<sup>1</sup> Contact author. E-mail: dlipson@cs.technion.ac.il

<sup>2</sup> CS Dept, The Technion and Molecular Diagnostics Department, Agilent Laboratories

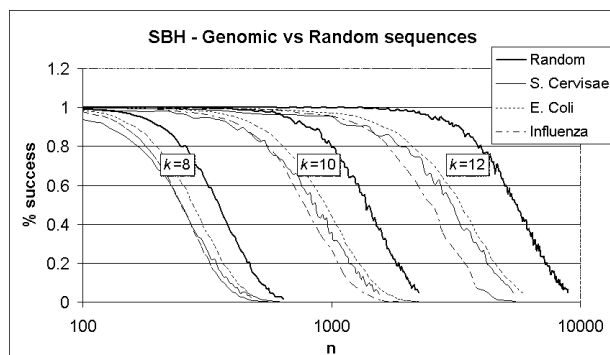


Figure 1: Rate of unique reconstruction for sequences of genomic origin, compared to sequences derived randomly.

How much of this difference is explained by the Markov structure of the genomic data? To better understand this issue we ran SBH simulations on data produced by Markov models of different orders, trained on the corresponding genomic data. The results confirm the hypothesis – even data generated by a Markov model of order 0 (the single base marginals) was sufficient to significantly lower the performance of SBH. Markov models of higher order yield even lower success rates, getting closer to the poor genomic data rates (Figure 2).

In summary, our results show a dramatic difference between simulated SBH performance on random data and genomic data. This difference can be partially explained by the repetitive nature of genomic data, a quality that is well captured in the sequence Markov structure. Our study therefore underlines and quantifies the need to use more accurate Markov models when evaluating applications of generic arrays by simulating assays on random sequences.

## References

- [1] R. Drmanac, I. Labat, I. Bruckner, and R. Crkevenjakov, Sequencing of Megabases Plus by Hybridization, 1989, *Genomics*, 4, pp 114-128.
- [2] Yu. P. Lysov, V.L. Florentiev, A.A. Khorlin, K.R. Khrapko, V.V. Shih, and A.D. Mirzabekov, 1998, Sequencing by Hybridization via Oligonucleotides, *Dokl. Acad. Sci. USSR*, Vol 303, pp 1508-1511.
- [3]: P.A. Pevzner, Yu. P. Lysov, K.R. Khrapko, A.V. Belyavsky, V.L. Florentiev and A.D. Mirzabekov, 1991, Improved Chips for Sequencing by Hybridization, *J of Biomol. Struct. and Dyn.*, Vol 9, No 2, pp 399-410.
- [4]: P.A. Pevzner, 1989, l-Tuple DNA Sequencing: Computer Analysis, *J of Biomol. Struct. and Dyn*, Vol 7, No 1, pp 63-73.
- [5] F.P. Preparata, A.M. Frieze, E. Upfal, 1999, On the Power of Universal Bases in Sequencing by Hybridization, in *Proceedings of RECOMB 1999*, ACM Press, pp 295-301.

When examining sequences derived from genomic databases, the success rate of the SBH algorithm is dramatically inferior to that observed for random data, as depicted in Figure 1. For genomic data we measured:  $n_{90}(8) = 140$ ,  $n_{90}(10) = 400$ ,  $n_{90}(12) = 1200$  – up to 3-fold worse than for random sequences. Difference in performance, between genomic and random, becomes more significant as  $k$  increases.

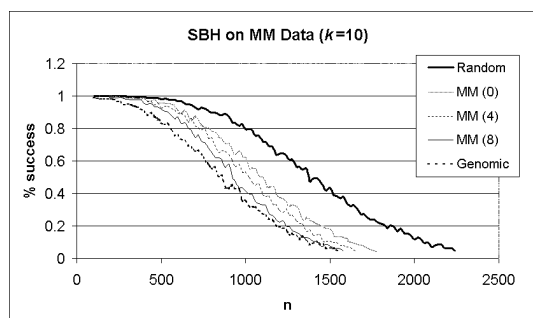


Figure 2: Rate of unique reconstruction for sequences created by Markov models of different orders, in comparison to randomly-generated and biological data.