

Improving Performance of DNA Sequencing-by-Hybridization with Gapped-Probes via Domain Knowledge

Zong He Cai and Hon Wai Leong
Department of Computer Science
National University of Singapore
Singapore 117543
{caizongh, leonghw}@comp.nus.edu.sg

Abstract

DNA sequencing-by-hybridization (SBH) was proposed as a powerful potential alternative to current sequencing methods by electrophoresis. Recent researches have shown that the use of SBH with gapped probes offers the potential for improved algorithmic sequence reconstruction. The algorithm, called *Adaptive Multi-Threshold (AMT)* algorithm, is able to recover long DNA sequence fragments (of length on the order of 10,000) and has performance that degrades gracefully with increasing hybridization noise. This paper presents an extension of AMTA that *incorporates sequence knowledge from a reference DNA family*. The extended algorithm, which we call the *Evaluative Multi-Threshold Algorithm (EMT)*, aims to improve the speed and robustness of the AMT algorithm by (a) penalizing non-feasible extensions and (b) early detection of spurious extension paths caused by noise – using domain knowledge from reference DNA sequences from the same family. To evaluate our Evaluative Multi-Threshold algorithm, we use both randomly generated DNA sequences as well as DNA sequences from several organisms (together with the mutated variants). On the generated DNA sequences, the results confirm the benefits of having the added domain knowledge in the reconstruction process – EMT outperforms AMT in terms of ?? and ??. When used with real DNA sequences, our research also exposes some weaknesses of existing SBH algorithms – the problem of repeats which cannot be disambiguated using short probes (gapped or otherwise). This is an important problem that must be overcome using other technological means for SBH to succeed.

Keywords:

Sequencing-by-hybridization, gapped probes, hybridization errors, domain knowledge.

1. Introduction

DNA sequencing-by-hybridization (SBH) was proposed over a decade ago [BS88, L88, D88, P89, P91, W95] as a potentially powerful alternatives to current electrophoresis techniques which are effective, but also costly and time consuming. SBH consists of two fundamental steps. The first is biochemical in nature and involves the acquisition of all sub-sequences (of a selected pattern) of a given *unknown target sequence* using the complementary hybridization process on a complete library of probes. This set of all sub-sequences is referred to as *sequence spectrum*. The second step, *combinatorial* in nature, is the *algorithmic reconstruction* of the original target sequence from its sequence spectrum. Although SBH is elegant in concept, serious difficulties in the fields of biochemistry (the use of a “universal” base in gapped probes) and combinatorial algorithms have prevented SBH from being operational [PJ04].

The issue of sequence noise is also important in sequence reconstruction algorithms and has an impact on the success rates and the target length of the reconstructed sequences. In the absence of sufficient data for realistic modelling of noise in DNA sequences, researchers have directed effort and attention to the design and implementation of SBH reconstruction algorithms which exhibit graceful degradation of efficiency on generated DNA sequences in presence of noise. Notable examples are the series of SBH reconstruction algorithms ([PFU99], [PU00], [PJ04], [LPSW02], [LPSW05]), the latest and most effective being the Adaptive Multi-Threshold algorithm [LPSW05]. This paper is a extension of these SBH algorithms.

With the availability of whole genome DNA sequences and advances in homology research¹, it is natural to ask if it is possible that domain knowledge could be applied to current reconstruction algorithms such as the Adaptive Multi-Threshold algorithm ([LPSW02, LPSW05]) to further enhance its robustness and performance under noisy conditions.

In our proposed Evaluative Multi-Threshold algorithm, it is hoped that the incorporation of domain knowledge would increase the robustness of the algorithm against hybridization noise and reduce reconstruction failures. This algorithm is also expected to provide speedy reconstructions of DNA sequences when supplied with similar sequences. The performance of our Evaluative Multi-Threshold algorithm would be evaluated against previous SBH algorithms using datasets from random sequence generators as well as real DNA sequences from the BLAST databases.

¹ The study of similarity in DNA or protein sequences between individuals of the same species or among different species.

2. Review of Adaptive-Threshold Algorithm

We briefly review the probing scheme and the Single Threshold (LPSW02) and Adaptive Multi-Threshold reconstruction algorithms (LPSW05). These algorithms form the basis of our own Evaluative Multi-Threshold Algorithm. We also review how the issue of noise in the spectrum is handled in these work – we adopt the same approach in this paper.

2.1. Noiseless Extension Algorithm

The standard reconstruction algorithm queries the spectrum for a set of feasible-extension probes given the current putative sequence, which is necessarily nonempty in the error-free case if the putative sequence is correct. If only one probe is returned, then we have a trivial one-symbol extension of the putative sequence viewed as a graph-theoretic path. Otherwise, we have an ambiguous branching and two or more competing paths are spawned; subsequently the algorithm attempts the breath-first extension of all paths issuing from the branching (and of all other paths spawned in turn by them) on the basis of the spectrum probes, unless it is found that all surviving paths have a common prefix, at which the prefix is concatenated to the putative sequence, and the process iterated.

2.2. Single Threshold Branching Algorithm

The Single Threshold Branching algorithm [LPSW02] is a breath-first extension of all 4 paths issuing from the branching (and of all other paths spawned in turn by them). Each path receives a score on the basis of the spectrum probes, a penalty score of 1 if the probe fails to exist in the spectrum and 0 otherwise. After normalisation, paths having scores above a threshold θ would be deemed to have accumulated excessive errors and pruned.

2.3. Adaptive Multi-Threshold (AMT) Algorithm

The AMT Algorithm [LPSW05] was developed to address the high computational cost of the Single Threshold Branching Algorithm. It presumes that majority of the sequence is re-constructible with the more efficient Noiseless Extension algorithm ($\theta=1$), given that there are few errors scattered among short stretches of symbols. A false negative causes an interruption to Noiseless Extension algorithm, either because none of the extension probes exists or a spurious path becomes extinct.

Under such circumstances, the algorithm assumes an error as occurred prior to the current extension. It steps back up the constructed sequence by J^* , an empirical value dependent on the target sequence length

m and the noise rate e , and sets $\theta=2$. The algorithm would then attempt local reconstruction by switching to the Single Threshold Branching Algorithm using this threshold until a unique sequence is built beyond the initial failure point, where it would reset $\theta=1$, and revert back to the Noiseless Extension algorithm. Failure in the local reconstruction would result in a repeat of the local reconstruction with an increased threshold, but no larger than the maximum threshold (a user-defined constant in our implementation).

2.4. Spectrum Noise

In the real world application of DNA sequencing, random hybridization errors (hybridization noise), in the form of false negatives (misses) and false positives (false-hits) occur. In the simulation of SBH, the approach to error control requires a presupposition of an error model, i.e. a formalization of the random process which produces hybridization errors.

Unfortunately, knowledge of the hybridization process is currently not available for precise quantification of the hybridization model and studies by researchers to the question of graceful degradation of the efficiency of the reconstruction process are based on the following error process (modified standard model):

1. Any spectrum probe can be suppressed with a fixed probability(false negatives);
2. Any probe at Hamming distance 1 from a correct spectrum can be added to the spectrum with a fixed probability(false positives);
3. Hybridization noise is expressed in terms of error rates for false negatives and positives.

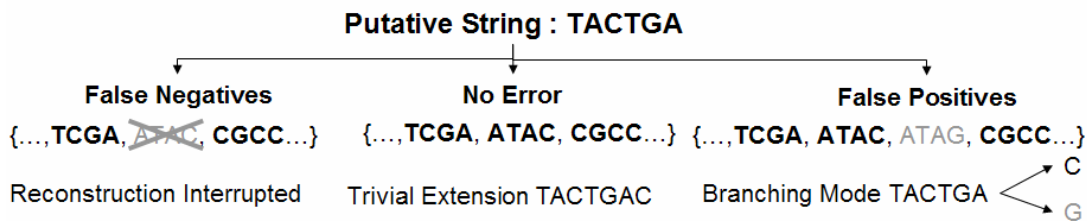


Figure 1 Effects of Noise on Probe Spectrum and Reconstruction Process

Intuitively, a false positive is much less detrimental to the reconstruction than a false negative. A false positive, merely adds one probe to the spectrum whereas the false negative irretrievably interrupts the extension process. False positive probes will only be accessed if they may act as feasible extensions (in the same way as fooling probes), and can be seen as simply increasing the pool of fooling probes, provided the false-positive rate remains reasonably small.

In the interest of simplicity and clarity, we assume that the false-positive rate is zero in this paper, and that the inclusion of false positives should only minutely complicate the reconstruction.

3. Our Evaluative Multi-Threshold Algorithm

The Evaluative Multi-Threshold algorithm is based on the fact that DNA sequences among different species do have many similarities when sequence fragments are considered, while certain patterns of base sequences just do not exist in real DNA. This homologous characteristic allows the possibility of supplying a reference data sequence as an extension to the Adaptive Multi-Threshold algorithm to assist in the SBH reconstruction during the error recovery phase.

3.1. Reference Spectrum Library

Our Evaluative Multi-Threshold Algorithm requires an additional spectrum to be generated from a reference DNA sample, using identical probes from the hybridization process of the original DNA sequence. The reference DNA is assumed to originate from a DNA library of known sequence, and hence the process of obtaining its spectrum is purely a computational step with no hybridization noise. This reference spectrum would then be further processed to generate a probability table that gives the reasonable occurrence probability for the extension of a putative string, as detailed below.

We define the Reference Spectrum Query Score to be the probability of an extension not occurring for a given query probe within the referenced domain spectrum. As an example, ‘A’ and ‘C’ are possible extensions for the query probe ‘ATG’, with ‘A’ occurring twice and ‘C’ eight times, while ‘T’ and ‘G’ do not exist in the spectrum. Hence the probability of ‘A’ and ‘C’ occurring are 20% and 80% respectively and the complement (1 – Probability of Occurrence) translates to the following scores.

Query Probe	ATG			
Extensions	A	C	T	G
Occurrence	2	8	0	0
% Occurrence	0.2	0.8	0.0	0.0
% Non Occurrence	0.8	0.2	1.0	1.0

Figure 2 Reference Spectrum Query Score

3.2. Weight

Weight is the factor which a score contributes towards the total score and is usually affected by the product between the weight and the relevant score. For the case of our algorithm, it represents the confidence in the level of similarity between the reference and sampled DNA.

3.3. Modified Scoring Function

Similar to the Adaptive Multi-Threshold algorithm, our Evaluative Multi-Threshold algorithm attempts to reconstruct the sequence with more efficient Noiseless Extension algorithm. Interruption to the Noiseless Extension algorithm would switch the algorithm to the Error Recovery algorithm previously described, but using a modified scoring function to determine the pruning of spurious paths.

The modified scoring first checks the sampled spectrum to determine if there is a possible extension. A negative could mean either an actual non-possibility of extension, or a probed that has been suppressed by hybridisation noise. The algorithm then checks the requested extension against the reference spectrum. In the case where the query from the reference spectrum reports true (i.e. probability of occurrence does not equal zero), there is a possibility that the probe in the original sampled spectrum has been suppressed, and the score returned would instead be a function of the value returned from the referenced spectrum table, and that of predefined weight. Otherwise, the score would be based on the sampled spectrum.

Modified Scoring Function for Evaluative Multi-Threshold Algorithm		Probe Result from Reference Spectrum	
		Extension exists	Extension does not exist
Probe Result from Sampled Spectrum	Extension exists	Score = 0	Score = 0
	Extension does not exist	Score = $1 - (Wt \cdot Prob)$	Score = 1

Figure 3 Summary of Modified Scoring Function

3.4. Evaluation Methodology

The performance of our Evaluative Multi-Threshold algorithm was evaluated the Noiseless Extension algorithm and Adaptive Multi-Threshold algorithm, using a variety of data sets of varying sizes (1000 ~ 16000 b.p). Each test was performed 200 times using data sets having similar characteristics.

Test No.	Test Description	Data Category	Reference Data (if applicable)	Source
1	Performance of various algorithms with random data in presence of noise	Random Data	Random Data	Sequence Generator
2	Performance of evaluative multi-threshold algorithm with mutated data against data from same species	Random Data with 10% point mutation	Random Data	Sequence Generator + Sequence Mutator
3	Performance of various algorithms with real data in presence of noise	Human X Chromosome Chimpanzee X Chromosome	Human X Chromosome Chimpanzee X Chromosome	FASTA DB
4	Performance of evaluative multi-threshold algorithm with mutated data against data from same species	Human X Chromosome with 10% point mutation Chimpanzee X Chromosome with 10% point mutation	Human X Chromosome Chimpanzee X Chromosome	FASTA DB + Sequence Mutator

Figure 4 List of Tests to be carried out

3.5. Results and Analysis

Performance of Noiseless Extension

From the graph, we see that the Noiseless Extension algorithm is unable to handle the reconstruction of the sampled DNA sequence if there is hybridization noise in all test cases.

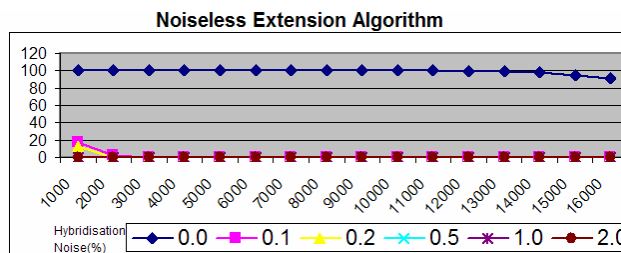


Figure 5 Performance Graph of Noiseless Extension

Algorithm Performance with Randomly Generated Sequences (Test 1)

When compared to the Noiseless Extension algorithm, both the Adaptive Multi-Threshold and our Evaluative Multi-Threshold algorithms display much better performance than the Noiseless Extension algorithm in the presence of noise, and exhibit graceful degradation as noise increases.

However, our Evaluative Multi-Threshold algorithm had the advantage of information from a reference sequence and showed better robustness to hybridisation noise, having an 80% success rate at 2% hybridisation noise with sequences of 12000 base pairs, compared to 5000 base pairs using the Adaptive Multi-Threshold Algorithm.

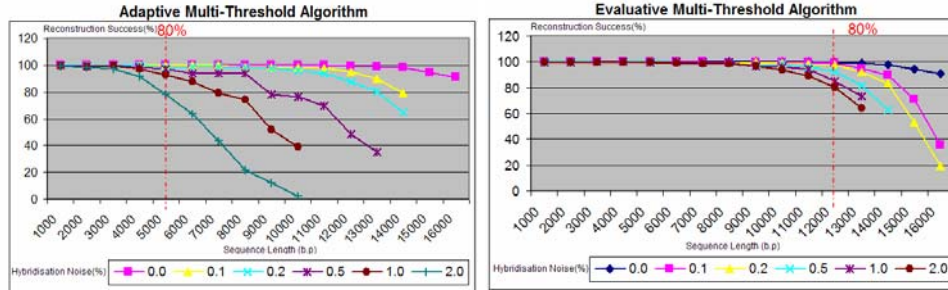


Figure 6 Performance Graphs of Adaptive Multi-Threshold algorithm and Evaluative Multi-Threshold algorithm for Randomly Generated DNA Sequences

Algorithm Performance with Randomly Generated Sequences (Mutated)

The 3 algorithms were evaluated again with sequences from the Homo sapiens X chromosome that has undergone a 10% point mutation. In the case of the Evaluative Multi-Threshold algorithm, the both the non-mutated and mutated corresponding sequences were used as reference to provide domain knowledge, with weights set at 50%.

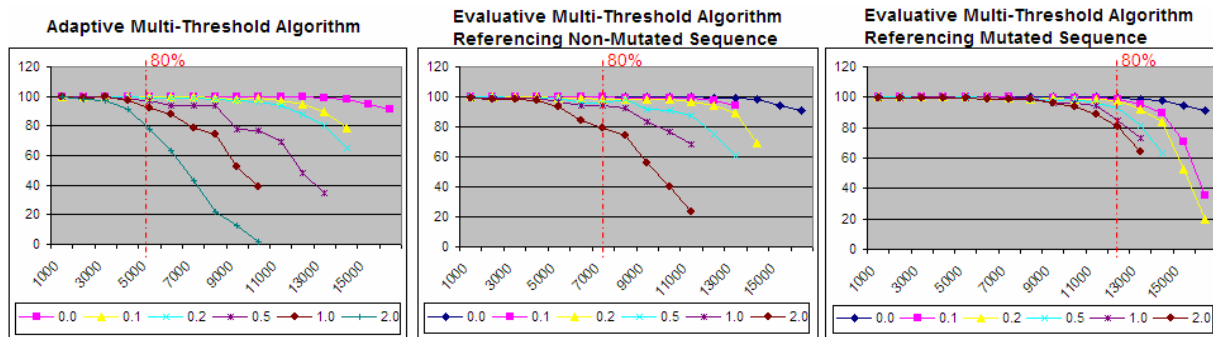


Figure 7 Performance Graphs for Mutated Randomly Generated DNA Sequences

As evident from the graphs, the Adaptive Multi-Threshold and the Evaluative Multi-Threshold algorithms perform very similarly to the reconstruction of the non-mutated sequence discussed above. The Evaluative Multi-Threshold algorithm referencing the non-mutated sequences fills in the performance gap, and shows about 80% reliability for reconstructing sequences of lengths up to 7500.

Algorithm Performance with Sequences from real DNA

All the algorithms quickly failed even on the shortest available DNA sequences across all species available, at all levels of noise. Investigation shows a high number of repeated patterns in the DNA, giving rise to fooling probes that prevent the reconstruction algorithm from definitively reconstructing the correct path.

With this behaviour, it was impossible to evaluate the performance of the various algorithms with respect to real DNA, but nonetheless, it highlighted a potential weakness dominant in the existing algorithms that should be looked into.

3.6. Conclusion

This research project has shown promise to the potential capabilities of the Evaluative Multi-Threshold algorithm, and has led to a better understanding to the strengths and weaknesses of the various algorithms considered. Once the issue of cycles existing real DNA sequences has been addressed, these existing reconstruction algorithms could be applied to real DNA to reap the benefits of DNA SBH under noisy conditions.

References

- [B97] Blazewicz J, Kaczmarek J, Kasprzak K, Markezicz WT, Weglarz J, DNA sequencing with positive and negative errors, *CABIOS* 13:151-158, 1997.
- [BS88] Baines W, Smith GC, A Novel method for DNA sequence determination, *J Theor Biol* 135:330-307, 1988
- [D88] Drmanac R, Labat I, Bruckner I, Crkvenjakov R, Sequencing of megabase plus DNA by hybridization, *Geneomics* 4:114-128,1988.
- [L88] Lysov YP, Florentiev VL, Kholin AA, Karapko KR, Shih VV, Mirzabekov AD. Sequencing by hybridization via oligonucleotides. A novel method, *Dokl Acad Sci USSR* 303:1508-1511, 1988.
- [LPSW02] Leong HW, Preparata FP, Sung WK, Willy H, On the Control of Hybridization Noise in DNA Sequencing-by-Hybridization, *WABI*, (2002), LNCS-2452, pp. 392-403.
- [LPSW05] Leong HW, Preparata FP, Sung WK, Willy H, Adaptive Control of Hybridization Noise in DNA SBH, *Journal of Bioinformatics and Computational Biology*, 3(1):1-20, 2005
- [P89] Pevzner PA, 1-tuple DNA sequencing : Computer analysis, *J Biomol Struct Dynamics* 7(1):63-73, 1989.
- [P91] Pevzner PA, Lysov YP, Khrapko KR, Belyavsky AV, Florentiev VL, Mirzabekov AD, Improved chips for sequencing by hybridization, *J Biomed Struct Dynamics* 9(2):399-410, 1991.
- [PJ04] Preparata FP, Oliver JS, DNA Sequencing by Hybridization Using Semi-Degenerate Bases, *Comput Biol* 11(4):753-765, 2004.
- [PU00] Preparata FP, Upfal E, Sequencing-by-hybridization at the information-theory bound: An optimal algorithm, *J Comput Biol* 7(3/4):621-630, 2000.
- [PFU99] Preparata FP, Fieze AM, Upfal E, On the power of universal bases in sequencing by hybridization, *Third Annual International Conference on Computational Molecular Biology*, April 11-14, Lyon, France, pp. 295-301, 1999.
- [W95] Waterman MS, *Introduction to Computational Biology*, Chapman and Hall 1995.