Available online at www.sciencedirect.com
ScienceDirect
Discrete Applied Mathematics 155 (2007) 857–867
DISCRETE APPLIED MATHEMATICS
www.elsevier.com/locate/dam

Restricting SBH ambiguity via restriction enzymes☆

Steven Skiena[a], Sagi Snir[b,*]

[a] *Department of Computer Science, SUNY Stony Brook, Stony Brook, NY 11794-4400*
[b] *Department of Mathematics, University of California, Berkeley, CA 94720, USA*

Received 15 July 2004; received in revised form 18 February 2005; accepted 19 February 2005
Available online 20 October 2006

Abstract

Sequencing by hybridization (SBH) is a proposed approach to DNA sequencing. The *SBH-spectrum* of the target sequence is a list of all k-mers occurring at least once in the sequence. Sequencing is successful if the SBH-spectrum is a result of only that sequence and ambiguous otherwise. Unfortunately, the expected number of sequences consistent with a given spectrum increases exponentially with the target sequence length.

In this paper, we extend previous work of [S. Snir, E. Yeger-Lotem, B. Chor, Z. Yakhini, SBH + RE—restriction enzymes dramatically enhance SBH, Technical Report, Department of Computer Science, The Technion, Haifa, Israel, 2002] to increase the resolving power of SBH by including information from enzymatic digestion assays. In addition to the hybridization assay, we conduct a small number of complete digestion assays using different restriction enzymes. The computational phase of identifying consistent sequences then combines the hybridization and digestion information. This combination of SBH and digestion assays significantly increases the length of sequences that can be uniquely determined. We give procedures for selecting the best enzymes for the job, prove that a variant of the reconstruction problem which includes an extra free parameter is hard, and give effective heuristics to improve search-based reconstruction algorithms. We also give a lower bound on the number of restriction enzymes required for unique reconstruction.

Keywords: Sequencing by hybridization; de Bruijn graphs; Restriction enzymes; NP-hardness; Information theory

1. Introduction

Sequencing by hybridization (SBH) [3,5,7,8,11,14] is a proposed approach to DNA sequencing where a set of single-stranded fragments (typically all possible 4^k oligonucleotides of length k) are attached to a substrate, forming a *sequencing chip*. A solution of single-stranded target DNA fragments are exposed to the chip. The resulting hybridization experiment gives the *SBH spectrum* of the target, namely a list of all k-mers occurring at least once in the sequence. This spectrum does not reveal the location of any k-mer in the sequence, nor does it count the number of its occurrences. A sequence is *SBH consistent* if it contains all and only k-mers from that spectrum. Sequencing is successful when only a single sequence is consistent with this spectrum, but *ambiguous* if multiple sequences are.

Unfortunately, the expected number of sequences consistent with a given spectrum increases exponentially with the sequence length. For example, the classical chip $C(8)$, with $4^8 = 65,536$ 8-mers suffices to reconstruct 200 nucleotide

☆ A preliminary version of these results appeared at the *Workshop on Algorithms in Bioinformatics* (WABI '02).

* Corresponding author.
E-mail addresses: skiena@cs.sunysb.edu (S. Skiena), ssagi@math.berkeley.edu (S. Snir).

0166-218X/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.dam.2005.02.034

long sequences in 94 of 100 cases [12] in error-free experiments. For $n = 900$, however, the expected number of consistent sequences rises to over 35,000. In this paper, we build on previous work [20] to increase the resolving power of SBH by including information from enzymatic digestion assays.

Several alternate approaches for increasing the resolving power of SBH have been proposed in recent years, including positional SBH [4], arrays with universal bases [16], interactive protocols [9,15], and tiling sequences with multiple arrays [18]. In contrast, the approach of [20] uses a very standard technique in molecular biology that predates oligonucleotide arrays by 20 years. *Restriction enzymes* identify a specific short recognition site in a DNA sequence, and cleave the DNA at all locations of this recognition site. In a complete digestion experiment, the product is a list of DNA fragment lengths, such that no fragment contains the recognition site. Measuring the length of these fragments yields a multiset of restriction fragment lengths called the *restriction enzymes* (*RE*) *spectrum*.

Snir et al. [20] propose the following procedure. In addition to the hybridization assay, they conduct a small number of complete digestion assays using different restriction enzymes and obtain restriction enzymes spectrums of the input sequence. The computational phase of identifying consistent sequences then combines the hybridization and digestion information. A sequence is (*SBH+RE*) *consistent* if it agrees with both the SBH and the RE spectrums. This combination of SBH and digestion assays significantly increases the length of sequences that can be uniquely determined.

In this paper, we study the power of combining SBH with restriction enzymes. Our results include:

- Although the idea of augmenting SBH with restriction enzymes is appealing, the algorithmic question of how to efficiently reconstruct sequences from such data remained open. In [20], a backtracking algorithm was proposed for reconstruction. In this paper, we show that a variant of the reconstruction problem, which has an extra free parameter, is NP-complete.
- We also propose additional search heuristics which significantly reduce the computation time. Such improvements are important, because for certain sequence lengths and enzyme set sizes the sequence is typically uniquely determined, yet naive backtracking cannot expect to find it within an acceptable amount of time.
- To gain more insight into the power of SBH plus restriction enzymes, we study the case of one digest from a theoretical perspective. This analysis shows when the restriction digest does and does not help uniquely determine the sequence from its SBH-spectrum.
- This analysis also suggests approaches for selecting restriction enzymes in response to the observed SBH-spectrum, so as to maximize the likelihood of unambiguous reconstruction. We give a heuristic to select the most informative restriction enzymes for a given SBH-spectrum.
- The resolving power of SBH plus restriction digests increases rapidly with the number of digests. We establish information-theoretic bounds on how many digests are necessary to augment the SBH-spectrum. We use insights from this analysis to select the cutter length and frequency for each digest to provide the optimal design of a sequencing experiment.

The rest of this paper is organized as follows: In Section 2 we give some background and definitions regarding SBH and RE and in Section 3 we prove the hardness of the SBH + RE problem and provide some useful heuristics to accelerate previous algorithms. Section 4 provides an insightful view on the structure of the problem through analytic inspection of a single RE digestion while Section 5 presents our method for selecting good REs based on the insights derived in the preceding section. In Section 6 we show the information theoretic bound and conclude in Section 7.

2. Background: SBH and restriction digests

As with most SBH algorithms, our work is based on finding postman walks in a subgraph of the de Bruijn digraph [6]. For a given alphabet Σ and length k, the *de Bruijn digraph* $G_k(\Sigma)$ contains $|\Sigma|^{k-1}$ vertices, each corresponding to a $(k-1)$-length string on Σ. As shown in Fig. 2, there will be an edge from vertex u to v labeled $\sigma \in \Sigma$ if the string associated with v consists of the last $k-2$ characters of u followed by σ. In any walk along the edges of this graph, the label of each vertex will represent the labels of the last $k-1$ edges traversed. Accordingly, each directed edge (u, v) of this graph represents a unique string of length k, defined by the label of u followed by the label of (u, v) (Fig. 1).

Pevzner's algorithm [13] interprets the results of a sequencing experiment as a subgraph of the de Bruijn graph, such that any Eulerian path corresponds to a possible sequence. As is typical (although regrettable) in SBH papers, here we assume error-free experimental data. When the sequence length is unknown, the reconstruction is unique iff the

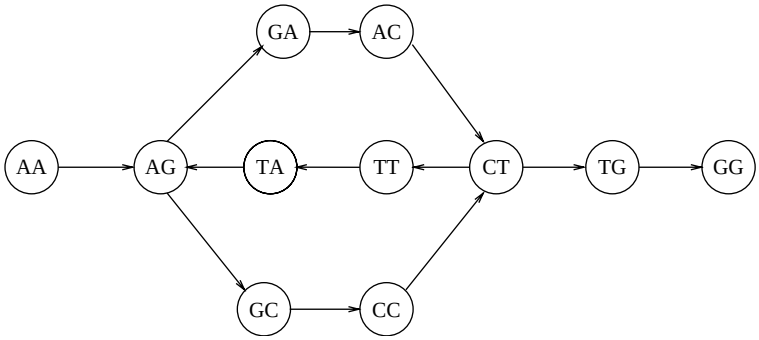

Fig. 1. The de Bruijn graph resulting from 3-mers of the sequence AGACTTAGCCTGG.

subgraph consists entirely of a directed induced path. When the sequence length is known, reconstruction is unique iff there is a unique postman walk of the appropriate length in the subgraph.

Restriction enzymes recognize and cut DNA molecules at particular patterns. For example, the enzyme *Eco*RI cuts at the pattern *GAATTC*. Enzymes are indigenous to specific bacteria, with the name of each enzyme denoting the order of discovery within the host organism (e.g. *Eco*RI was the first restriction enzyme discovered in *E. Coli*). Restriction enzymes cut double-stranded DNA, and are classified as either types I or II depending whether they cleave the strand at specific locations, with those that do (type II) significantly more important for biotechnology. Enzymes can be further classified according to whether they leave behind blunt or sticky ends, i.e. double- or single-stranded end fragments. Rebase [17] maintains a complete list of all known restriction enzymes, including cutter sequences, literature references, and commercial availability. As of July 12, 2004, 9384 different enzymes were known, defining at least 255 distinct cutter sequences. Cutter sequence lengths range in length from 2 to 15 bases. Although most enzymes cut at specific oligonucleotide base patterns, other enzymes recognize multiple sequences by allowing variants at specific base positions. For example, the cutter AACNNNNNNGTGC matches from the $5'$ to $3'$ end any sequence starting AAC, ending GTGC where they are separated any sequence of exactly six bases. The output of a restriction assay is the RE spectrum which is the multiset of RE fragment lengths induced by the restriction enzyme.

In this paper, we will limit our attention to cutter sequences without wildcard bases. We assume that the DNA sequence is cleaved at the start (i.e. position 0) of the restriction site. As in [20], we assume that all possible cutter sequences of a given length have associated enzymes. While this assumption is not true (only 17 distinct 4-cutters and 85 distinct 6-cutters currently appear in Rebase), we do not believe this materially changes our results.

3. Complexity of reconstruction and search heuristics

The algorithmic problem of reconstructing SBH data augmented with restriction digests is not as simple as that of pure SBH, however. Snir et al. proposed a backtracking algorithm to enumerate all possible sequences consistent with the data. This algorithm backtracks whenever a prefix sequence violated some property of the data, such as length, oligonucleotide content, or position of restriction sites, and is described in detail in [20]. We define the *sequence reconstruction from SBH plus restriction digests* problem as:

Input: An SBH spectrum S for the array of all k-mers, a set of e partitions of integer n, each corresponding to the results of a restriction digest with a particular cutter sequence.

Output: Does there exist a sequence T which is consistent with both the SBH spectrum S and the set of restriction digest results?

Snir et al. [20] provided no complexity analysis for their algorithm and since the number of postman walks in the de Bruijn graph can be exponential in the length of the sequence, this is not an efficient algorithm. Therefore, the question of whether there exists a provably efficient reconstruction algorithm for the SBH + RE problem remained open. We show that a generalization of this problem (relaxing the constraint of a DNA-size alphabet) is NP-complete.

Theorem 1. *Sequence reconstruction from SBH plus restriction digests with an unlimited alphabet is* NP*-complete even with just one restriction digest.*

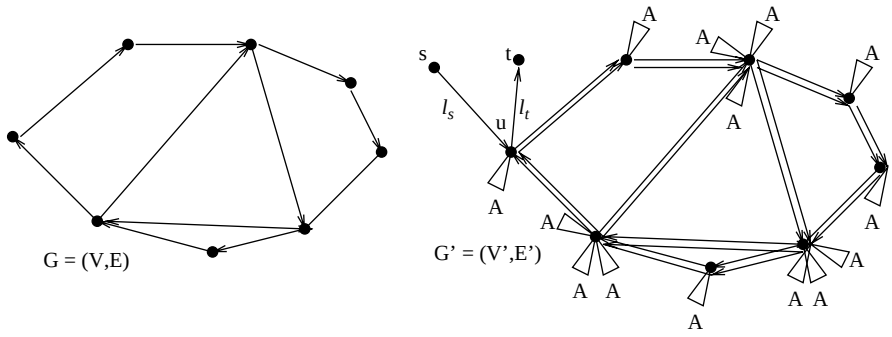

Fig. 2. Reducing Hamiltonian Cycle in G to SBH + RE in G'.

Proof. We show a reduction from Hamiltonian cycle in directed Eulerian graphs [10], by constructing a set of strings returned from an SBH-spectrum, and an associated set of restriction digest lengths such that reconstructing a consistent sequence involves solving the Hamiltonian cycle problem.

Our reduction from input graph $G = (V, E)$ is illustrated in Fig. 2. We use a large alphabet Σ, and spectrum length $k = 2$. There will be a distinct letter in Σ for each vertex in V, with each directed edge $(x, y) \in E$ represented by the string xy. We then double all the edges of G, turning it into a multigraph. Note that this has no impact on whether G contains a Hamiltonian cycle. We then augment this graph with the following structures to construct a new graph $G' = (V', E')$:

- Starting and ending disjoint paths from two new vertices s and t, of length ℓ_s and ℓ_t, respectively, joining the original graph at an arbitrary vertex u. Since these paths are disjoint, all vertices in them are new unique vertices. These define the beginning and end of the sequence.
- We add a sequence A, composed of new unique vertices and connected from and to all of the original vertices $v \in V$, forming a loop. (For clarity of representation, the loops in Fig. 2 appear disjoint, however, they all pass through the same vertices of A, but in different, parallel edges.) For every $v \in V$ there are $d(v) - 1$ such loops incident on v, where $d(v)$ is the out-degree of v in the doubled graph. Vertex u is given an additional such loop.

In addition, we provide the results of the restriction assay with A as the cutter, yielding the multiset of distances $l_s + |V| + 1$, $l_t + |A| + 2$, and $2|E| - |V| - 1$ fragments of length $|A| + 2$. To simplify matters, we demand that these lengths be distinct (e.g. $|V| + 1$, $4|V| + 1$ and $5|V| + 1$).

In other words, on an input graph G as an instance to the Hamiltonian Cycle problem, the reduction outputs a SBH spectrum corresponding to the de Bruijn graph G' and a RE spectrum with lengths as above. □

Claim 1. *A sequence is consistent with the de Bruijn graph G' and the restriction assay if and only if it describes a Hamiltonian cycle in G.*

Proof. $\Longrightarrow$ Let C be a Hamiltonian cycle in G. We construct the following Eulerian tour in G' that will induce the required restriction assay: take the path $s \to u$ and then the cycle C and the A loop from and to u. This gives the RE fragment $\ell_s + |V| + 1$. Since G is Eulerian, the graph of uncovered edges after covering the edges of C is still Eulerian and also connected, and therefore we can cover all uncovered edges of G' with the tour to the special sequence A interleaved. Hence, we get all the length-$(|A| + 2)$ RE fragments. We conclude with taking the path $u \to t$ to obtain the $\ell_t + |A| + 2$ fragment.

$\Longleftarrow$ We now prove that an SBH+RE consistent sequence defines a Hamiltonian cycle in G. Remember that our set of RE fragments is $\ell_s + |V| + 1$, $\ell_t + |A| + 2$, and $2|E| - |V| - 1$ fragments of length $|A| + 2$. By our definition, the walk must start at vertex s, end at vertex t and the sequence is cut at every entrance to the A-loop. Seeking for contradiction, suppose the input graph does not include a Hamiltonian cycle. Then in order to satisfy a RE fragment of length $\ell_s + |V| + 1$ some vertex *must* be visited at least twice. We distinguish between the original graph edges and A-loop edges. In the sequel, an edge that was already traversed is removed from the graph.

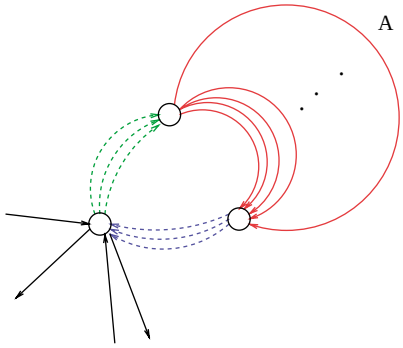

Fig. 3. In a doubly visited vertex, the in+out degree of original (solid) edges is smaller than the in+out A-loop (dashed) edges.

So after collecting the $(\ell_s + |V| + 1)$th RE fragment, the next RE fragments that have to be satisfied are the $2|E| - |V| - 1$ fragments of length $|A| + 2$. It is true that taking the A-loop detour does not force returning to the same vertex, however, focusing on a doubly visited vertex (Fig. 3) reveals the following observations:

Observation 1. If a vertex v was visited more than once in the first part of the walk, there is an imbalance between its graph in+out degree and its A-loop in+out degree.

Denote a vertex whose graph in+out degree is greater than its A-loop in+out degree, an *imbalanced* vertex and otherwise *balanced*. Now, since eventually, every vertex needs to be balanced (i.e. degree 0 in both edge types), we need to fix this imbalance.

Observation 2. The only way to restore the balance between the graph in+out degree and the A-loop in+out degree is to take two or more successive A-loops at the imbalanced vertex v.

Corollary 1. *In order to cover all edges incident at an imbalanced vertex, there must be a* $|A + 1|$*-long RE fragment.*

However, since our RE spectrum does not contain a RE fragment of length $|A + 1|$, the string defined by this tour is not RE consistent—contradiction to the assumption. □

3.1. Heuristics for accelerating the search algorithm

Beyond this negative result, we improved the search algorithm of [20] with two new pruning criteria which yield better performance. Before showing these improvements, we briefly describe the algorithm of [20]. Assuming a short prefix and suffix of the target sequence (say, $2k$) are known, the algorithm searches all postman walks from the prefix in the de Bruijn graph. Along a postman walk P, all k-mers and RE fragments encountered so far are recorded. The algorithm abandons P (i.e. backtracks) when

- the total length of the uncovered paths plus the length of the current walk exceeds the sequence length (with calibration of the overlaps).
- The multiset of RE fragments encountered along P cannot be completed to the target RE spectrum.

If P has reached the required length, and the suffix of the sequence defined by P is the required one, then the sequence defined by P is a SBH + RE consistent sequence.

We now describe the proposed improvements:

- *Improved sequence length cutoffs*: The results of restriction digests implicitly tell of the length of the target sequence. Thus, we are interested in finding all SBH-consistent sequences of a given length. Such sequences correspond to a given postman walk on the appropriate de Bruijn subgraph.

The state of the backtracking computation is represented by the prefix of the putative sequence under construction. Snir et al. observed that we can backtrack whenever the length of this prefix plus the length of all thus-far unvisited (uncovered) edges of the graph exceeds that of the target. However, this is an underestimate of the missing length: Let v be a vertex in G_F with u_i uncovered in-edges and u_o uncovered out-edges, at some configuration c on a path p to consistent sequence. Then p must also traverse at least $|u_i - u_o|$ times covered edges.
This gives rise to the following pruning technique: A vertex v with u_i uncovered in-edges and u_o uncovered out-edges is *in-imbalanced* if $u_i > u_o$ and *out-imbalanced* if $u_i < u_o$. Let v be an in-imbalanced vertex and let e_m be v's out-going edge (covered or uncovered) with minimal length l_m. Then we can bound the missing length by at least $(u_i - u_o)(l_m - k + 1)$ and potentially backtrack earlier. This claim is true for all vertices of G_F, however we must take precaution not to double-count the same edge by two adjacent vertices. To prevent this, we separate the extra length implied by in-imbalanced vertices from that implied by out-imbalanced vertices, and add the bigger of these quantities to be the missing length.

- *Strong connectivity cutoffs*: The strongly connected components of a digraph are the maximal subgraphs such that every pair of vertices in a component are mutually reachable. Partitioning any digraph into strongly connected components leaves a directed acyclic graph of components such that it is impossible to return to a parent component by a directed path.
This gives rise to the following pruning principle. Let $G_r = (V, E_r)$ be the residual graph of uncovered edges. Let C_1 and C_2 be two different strongly-connected components in G_r, and edge $(u \to v) \in E_r$ link the two components, i.e. $u \in C_1$ and $v \in C_2$. Then we can backtrack on any prefix which traverses edge (u, v) before covering all edges in C_1.

Each of these two techniques reduced search time by roughly 70% on typical instances. When operated in conjunction, the search time was typically reduced by about 80%. All the algorithms were implemented in Java, and run on Pentium 3 PCs.

4. Understanding single digests

Restriction digest data usually reduces the ambiguity resulting from a given SBH-spectrum, but not always. We can better understand the potential power of restriction digests by looking at the topology of the given de Bruijn subgraph.

We define the notion of a partially colored graph to integrate the information from an SBH-digest with a given collection of RE-digests. A graph $G(V, E)$ is *partially colored* if a subset of vertices $V' \subset V$ are assigned colors, and the vertices $V - V'$ remain uncolored. Let $G = (V, E)$ be the subgraph of the de Bruijn graph of order $k - 1$ defined by a given SBH-spectrum S, and R be a set of strings $\{r_1, \ldots, r_c\}$. We say that G is *partially colored with respect to* R iff the coloring of a given $v \in V$ implies that there exists a string $r \in R$ where the $|r|$-prefix of the $k - 1$-mer associated with v equals r. That is, a colored vertex represents a restriction site. We assume that $r \leqslant k - 1$, as will naturally be the case in reasonable problem instances.

For certain partially colored graphs, the restriction digest data is sufficient to unambiguously reconstruct sequences not completely defined by the SBH-spectrum alone. Fig. 4 depicts such a graph. Denote a colored vertex (restriction site) by a filled circle. Without a restriction digest, the postman walk *abcde* will be SBH-consistent with *adbce*. However, since the restriction digests of the two sequences are $\{|ab|, |cde|\}$ and $\{|adb|, |ce|\}$, respectively, such a digest will be sufficient to disambiguate them provided $|ce| \neq |ab|$.

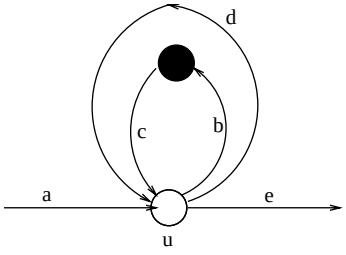

Fig. 4. A partially colored de Bruijn graph which might be disambiguated on the basis of restriction digest data.

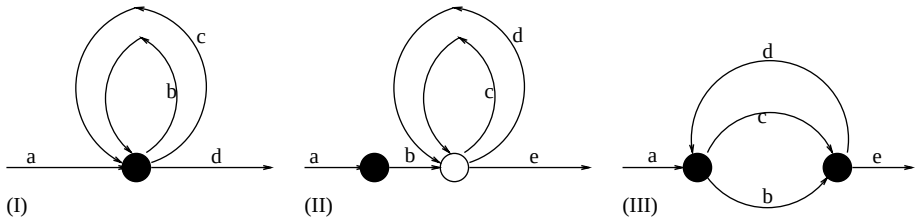

Fig. 5. Three hopeless digraphs. For every postman walk in G, there exists another postman walk with the same length and same set of distances between colored vertices.

Not all partially colored graphs G have this property. We say that G is *hopeless* with respect to its partial coloring if every postman walk P on G has another postman walk P' on G such that $|P| = |P'|$ and the multisets of distances between successive colored vertices along P and P' will be the same.

Fig. 5 depicts three cases of hopeless graphs. The graph in Fig. 5(I) is topologically the same as in Fig. 4, but now the colored vertex is the junction of two loops. The order of traversal of the two loops cannot be distinguished by the RE-spectrum. The graph in Fig. 5(II) is hopeless because the cut at u cannot eliminate the ambiguity from elsewhere in the graph. Finally, the graph in Fig. 5(III) is hopeless since every postman walk must traverse paths c and b in some order. Reversing this order (by a tandem repeat) causes no change to either the SBH or RE-spectrums.

We now consider the problem of characterizing sequences which are uniquely defined by SBH plus restriction digests. Conditions under which a sequence is uniquely defined by its SBH-spectrum were established by Pevzner and Ukkonen and formalized in [2]:

Theorem 2. *A word W has another word W' with the same SBH-spectrum iff the sequence of overlapping $k-1$-mers comprising W denoted $\overrightarrow{W}$ has one of the forms*:

- $\alpha a \beta a \gamma a \delta$.
- $\alpha a \beta b \gamma a \delta b \varepsilon$.

where $a, b \in \Sigma^{k-1}$ and $\alpha, \beta, \gamma, \delta, \varepsilon \in (\Sigma^{k-1})^*$.

As demonstrated by the figures above, a single restriction digest R can resolve ambiguities in the SBH-spectrum S. We make this intuition formal below. We say that a $(k-1)$-mer is *colored* if it defines a colored vertex in the partially colored de Bruijn graph of S plus R.

Theorem 3. *A word W with k-mer representation $\overrightarrow{W}$ which is uniquely defined by its SBH-spectrum and a single restriction digest satisfies all of the following properties*.

(1) *$\overrightarrow{W}$ does not contain a colored $k-1$-mer a such that $\overrightarrow{W} = \alpha a \beta a \gamma a \delta$, i.e. a does not occur 3 times in W (case I in Fig. 5).*
(2) *$\overrightarrow{W}$ does not contain two colored $k-1$-mers a and b such that $\overrightarrow{W} = \alpha a \beta b \gamma a \delta b \varepsilon$ (case III in Fig. 5).*
(3) *$\overrightarrow{W}$ does not contain a substring $\overrightarrow{W'}$ consisting entirely of uncolored $k-1$-mers, where $\overrightarrow{W'} = \alpha a \beta a \gamma a \delta$ or $W' = \alpha a \beta b \gamma a \delta b \varepsilon$ (case II in Fig. 5).*

Proof. We analyze each case separately. For a word W' and a restriction enzyme r, we denote by $\rho(W')$ the set of RE fragments obtained by applying r on W'. For case 1, let W_1, W_2, W_3, W_4 such that $\overrightarrow{W_1} = \alpha$; $\overrightarrow{W_2} = a, \beta$; $\overrightarrow{W_3} = a, \gamma$ and $\overrightarrow{W_4} = a, \delta$. Since a is colored then $\rho(W) = \rho(W_1) \cup \rho(W_2) \cup \rho(W_3) \cup \rho(W_4)$. Now let W^* such that $\overrightarrow{W^*} = \alpha a \gamma a \beta a \delta$. Then by Theorem 2 W and W^* are SBH-consistent, and $\rho(W^*) = \rho(W_1) \cup \rho(W_3) \cup \rho(W_3) \cup \rho(W_4) = \rho(W)$—contradiction to the fact that W is uniquely defined.

For case 2, let W_1, W_2, W_3, W_4, W_5 such that $\overrightarrow{W_1} = \alpha$; $\overrightarrow{W_2} = a, \beta$; $\overrightarrow{W_3} = a, \gamma$; $\overrightarrow{W_4} = a, \delta$ and $\overrightarrow{W_5} = a, \varepsilon$. Since a is colored then $\rho(W) = \rho(W_1) \cup \rho(W_2) \cup \rho(W_3) \cup \rho(W_4 \cup \rho(W_5)$. Now let W^* such that $\overrightarrow{W^*} = \alpha a \delta b \gamma a \beta b \varepsilon$. Then by Theorem 2

W and W^* are SBH-consistent, and $\rho(W^*) = \rho(W_1) \cup \rho(W_3) \cup \rho(W_3) \cup \rho(W_4) \cup \rho(W_5) = \rho(W)$—contradiction to the fact that W is uniquely defined.

Case 3 follows by arguments analogous to those of the previous two cases. □

Theorem 4. *Let $i_1, \ldots, i_d$ represent the position of all interesting positions in a word W, where a position is interesting if*; (1) *it corresponds to a colored vertex*, (2) *it corresponds to a vertex which appears three or more times*, *or* (3) *it corresponds to a vertex of a tandem repeat in W. Then W is uniquely defined by its SBH-spectrum and a single restriction digest if it satisfies all of the above conditions and no two subsets of $\bigcup_{j=1}^{d+1} i_j - i_{j-1}$ sum to the same value*, *where $i_0 = 0$ and $i_{d+1} = n$* (*the sequence length*).

Proof. Let f_j be the fragment defined between two consecutive interesting points i_j and i_{j+1}. It is clear that every SBH-consistent sequence is a permutation of the fragments between points of type (2) or (3), and such a permutation cannot yield a new fragment. Now, by condition (3) of Theorem 3, every triple or tandem repeat contains at least one restriction site. Thus, every shuffling of fragments yields a new set of fragments between restriction sites, and by the assumption, this new set has total length not existing in the original sequence. □

5. Selecting enzymes to maximize resolution

Observe that there is nothing in the SBH + RE protocol of [20] which requires that the experiments be done in parallel. This means that we could wait for the SBH-spectrum of the target sequence and use this information to select the restriction enzymes which can be expected to most effectively disambiguate the spectrum.

Based on the observations of Theorem 3, we note that digests which color high-degree vertices inherently leave ambiguities. Further, digests which do not include a restriction site breaking regions between tandem repeats or sequence triples cannot resolve the alternate sequences described in Theorem 2.

These observations suggest the following heuristic to select good enzymes. Randomly sample a number of appropriate length sequences consistent with the observed SBH-spectrum (note that these are only SBH consistent, not SBH + RE consistent). Simulate a restriction digest with each possible cutter sequence on each sampled sequence. For each, count the number of forbidden structures which lead to ambiguous reconstructions of the sampled sequences. Select the enzymes leading to the smallest number of such structures.

The forbidden structures (or *violations*) we seek to avoid in sequence s for enzyme cutter sequence e are:

- Tandem repeat $\alpha a \beta b \gamma a \delta b \varepsilon$, where e is the prefix of a and b, or e does not cut $a\beta b\gamma a\delta b$.
- Triple repeat $\alpha a \beta a \gamma a \delta$, where e is the prefix of a or e does not cut $a\beta a\gamma a$.

We define a violation with respect to sequence s and cutter-sequence e for every one of these forbidden structures. We sort the enzymes first according to the total number of violations, breaking ties based on the maximum number of violations. We select the first k enzymes in this sorted list for use in the search algorithm. The distribution of violations per enzyme for sample sequences of length $n = 1000$ and $n = 1500$ are shown in Fig. 6.

Generating random SBH-consistent sequences of the given length is a non-trivial problem. Indeed, [19] proves the problem of generating an SBH-consistent sequence of a given length is NP-complete, although hardness requires very long sequences (i.e. the problem is not strongly NP-complete). The complexity issue can be readily worked around by heuristics. We employ a search-based algorithm to find a suitable sequence. Once we have a starting sequence, we can interchange fragments at triple subsequences and tandem repeats to generate other sequences by the arguments of Theorem 2.

6. Selecting cutter length and frequency

In this section, we analyze the impact of cutter length and number of digests on the performance of SBH + RE to suggest the best design for such an experiment. To do so, we use tools from the theory of integer partitions [1]. We observe that each complete digest (including multiplicities) returns a partition of the sequence length n. Since, on average, an r-cutter cuts every 4^r bases, the expected number of parts resulting from such a digest is $n/4^r$. We seek to get the maximum amount of information from each restriction digest. As illustrated in Fig. 7, the number of

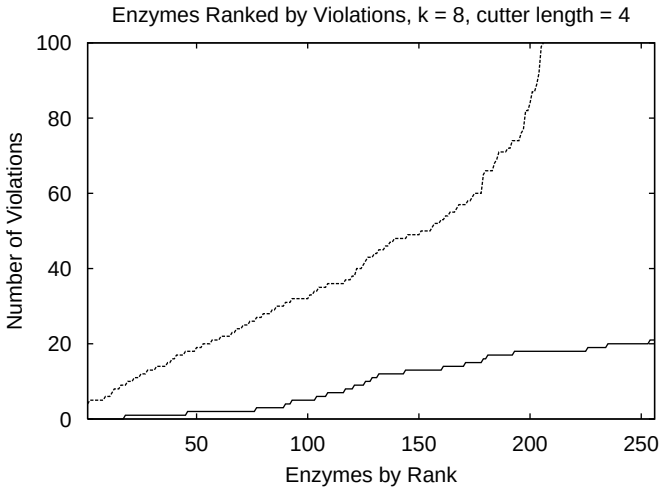

Fig. 6. Number of violations per enzyme in typical random sequences of $n = 1000$ and $n = 1500$.

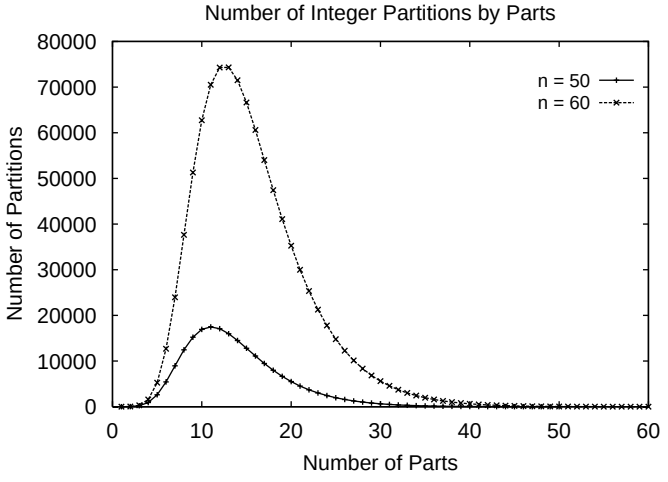

Fig. 7. Number of partitions by Parts for $n = 50$ and $n = 60$.

partitions of n with p parts peaks around $p = 2\sqrt{n}$, so the ideal cutter-length r will yield this many parts. This occurs at $r = (\log_2 n)/4 - 1$, a function growing slowly enough that it remains less than $r = 2.4$ for $n \leqslant 10,000$.

We can use similar arguments to obtain a lower bound on the number of digests needed as a function of the length of the sequence:

Observation 3. Let S be a random sequence over a four letter alphabet. The expected number of restriction digests needed to disambiguate the k-mer SBH-spectrum of S is at least D, where

$$D \geqslant n^{3.5}/(24(\lg e)(\sqrt{2/3})(4^{k-1})^2).$$

Our argument is as follows. Consider the SBH spectrum associated with all k-mers of a sequence S of length n. The probability $P(k, n)$ that any given k-mer occurs more than once in S may be calculated as

$$P(k,n) \approx \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(\frac{1}{4}^{k}\right)^2 \approx \left(\frac{1}{2}\right)(n/4^k)^2.$$

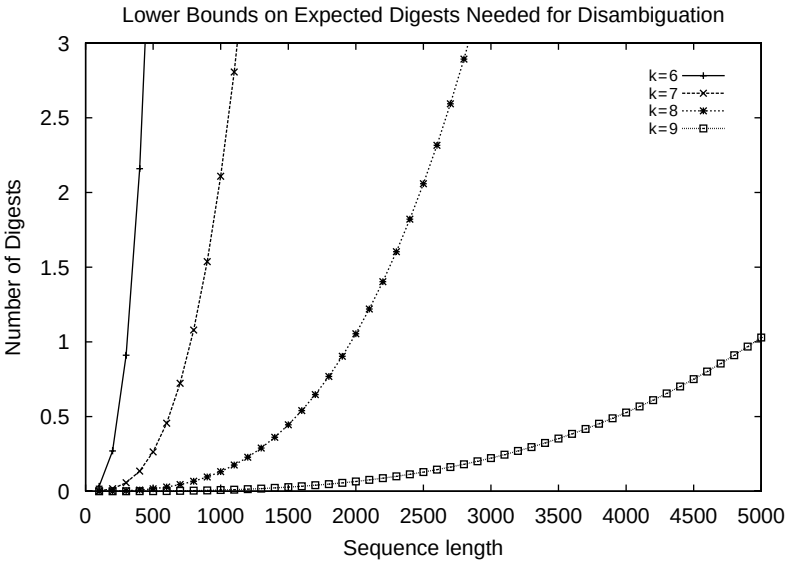

Fig. 8. Lower bounds on expected number of digests required as a function of sequence and k-mer length.

Thus, the expected number of vertices of out-degree $\geqslant 2$ in the resulting de Bruijn subgraph is

$$v \approx 4^{k-1} \times P(k-1, n) \approx n^2/(2 \cdot 4^{k-1}).$$

Given that we have v vertices of out-degree greater than 2, we can compute a bound on number of postman paths satisfying the spectrum. Whenever a tandem repeat $a, \ldots, b, \ldots, a, \ldots, b$ occurs in S, the two subpaths can be shuffled, creating sequence ambiguity. Thus the number of paths is at least 2^t, where t is the number of tandem repeats.

The probability that two out-degree 2 vertices create a tandem repeat between them is $\frac{1}{3}$, since there are six possible orderings of the four sites, two of which are tandem. Thus v high degree vertices gives rise to an expected $\approx 2^{v^2/6}$ paths.

The output of each restriction digest is an integer partition of n describing the number and length of the fragments. The number of integer partitions of n is asymptotically:

$$a(n) \approx (1/4n)(1/\sqrt{3})e^{(\pi\sqrt{2n/3})}$$

as $n \to \infty$, by Hardy and Ramanujan [1]. Thus the information content of a restriction digest is $\lg(a(n)) \approx (\lg e)(\sqrt{2n/3})$ bits.

We need at least enough bits from the digests to distinguish between the $\approx 2^{v^2/6}$ paths, i.e. the binary logarithm of this number. Since $v \approx n^2/(2 \cdot 4^{k-1})$, we need approximately $n^4/(24 \cdot 4^{2(k-1)})$ bits. Therefore the number of digests D is $D \geqslant n^{3.5}/(24(\lg e)(\sqrt{2/3})(4^{k-1})^2)$.

Despite the coarseness of this bound (e.g. ignoring all sequence ambiguities except tandem repeats, and assuming each partition returns a random integer partition instead of one biased by expected number of parts) it does a nice job matching our experimental data. Note that the bound holds for smart enzyme selection as well as random selection. Fig. 8 presents this lower bound for $6 \leqslant k \leqslant 9$ over a wide range of sequence lengths. In all cases, the expected number of enzymes begins to rise quickly around the lengths where sequence ambiguity starts to grow.

7. Concluding remarks and future research directions

We studied the new technique of [20] that is based on combining data from two biological sources. We showed that by intelligent selection of restriction enzymes, the probability of obtaining a unique SBH + RE consistent sequence increases.

Certain research directions suggest themselves for further investigation. In this work we dealt solely with error free model. It would be interesting to extend the results shown here to a more realistic model which takes into account noisy data. In this respect, there are two types of noises: noisy hybridization data and noisy restriction data. Handling both types of errors appears to be challenging.

Acknowledgments

We wish to thank Zohar Yakhini for many enlightening discussion on this contribution and in particular for his help in the proof of Theorem 1.

References

[1] G. Andrews, The Theory of Partitions, Addison-Wesley, Reading, MA, 1976.
[2] R. Arratia, D. Martin, G. Reinert, M. Waterman, Poisson process approximation for sequence repeats, and sequencing by hybridization, J. Comput. Biol. 3 (1996) 425–463.
[3] W. Bains, G. Smith, A novel method for nucleic acid sequence determination, J. Theor. Biol. 135 (1988) 303–307.
[4] A. Ben-Dor, I. Pe'er, R. Shamir, R. Sharan, On the complexity of positional sequencing by hybridization, Lecture Notes in Computer Science, vol. 1645, 1999, pp. 88–98.
[5] A. Chetverin, F. Kramer, Oligonucleotide arrays: new concepts and possibilities, Bio/Technol. 12 (1994) 1093–1099.
[6] N.G. de Bruijn, A combinatorial problem, Proc. Kon. Ned. Akad. Wetensch 49 (1946) 758–764.
[7] R. Dramanac, R. Crkvenjakov, DNA sequencing by hybridization, Yugoslav Patent Application 570, 1987.
[8] S. Fodor, J. Read, M. Pirrung, L. Stryer, A. Lu, D. Solas, Light-directed, spatially addressable parallel chemical synthesis, Science 251 (1991) 767–773.
[9] A. Frieze, B. Halldorsson, Optimal sequencing by hybridization in rounds, in: Proceedings of Fifth Conference on Computational Molecular Biology (RECOMB-01), 2001, pp. 141–148.
[10] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-completeness, W. H. Freeman, San Francisco, 1979.
[11] Y. Lysov, V. Florentiev, A. Khorlin, K. Khrapko, V. Shik, A. Mirzabekov, Determination of the nucleotide sequence of dna using hybridization to oligonucleotides, Dokl. Acad. Sci. USSR 303 (1988) 1508–1511.
[12] P. Pevzner, Y. Lysov, K. Khrapko, A. Belyavski, V. Florentiev, A. Mizabelkov, Improved chips for sequencing by hybridization, J. Biomol. Structure Dynamics 9 (1991) 399–410.
[13] P.A. Pevzner, l-tuple DNA sequencing: computer analysis, J. Biomol. Structure Dynamics 7 (1989) 63–73.
[14] P.A. Pevzner, R.J. Lipshutz, Towards DNA sequencing chips, 19th International Conference on Mathematical Foundations of Computer Science, vol. 841, 1994, pp. 143–158.
[15] V. Phan, S. Skiena, Dealing with errors in interactive sequencing by hybridization, Bioinformatics 17 (2001) 862–870.
[16] F.P. Preparata, E. Upfal, Sequencing-by-hybridization at the information-theory bound: an optimal algorithm, in: Proceedings of Fourth Conference on Computational Molecular Biology (RECOMB-00), 2000, pp. 245–253.
[17] R. Roberts, Rebase: the restriction enzyme database. ⟨http://rebase.neb.com⟩, 2001.
[18] R. Shamir, D. Tsur, Large scale sequencing by hybridization, in: Proceedings of Fifth International Conference on Computational Molecular Biology (RECOMB-01), 2001, pp. 269–277.
[19] S. Skiena, G. Sundaram, Reconstructing strings from substrings, J. Comput. Biol. 2 (1995) 333–353.
[20] S. Snir, E. Yeger-Lotem, B. Chor, Z. Yakhini, SBH + RE—restriction enzymes dramatically enhance SBH, Technical Report, Department of Computer Science, The Technion, Haifa, Israel, 2002.