DNA Assembly From Paired Reads as 2-D Jigsaw Puzzles

Eren Şaşoğlu and David Tse
University of California, Berkeley
Berkeley, CA 94720, USA
{eren,dtse}@eecs.berkeley.edu

***Abstract*—We study the information theoretic limits of DNA assembly from paired reads. Each paired read consists of two subsequences of the DNA separated by a certain genomic distance. We show that this problem can be naturally cast as assembly of 2-D jigsaw puzzles. Using this representation, a necessary condition for assembly is derived and is shown to be nearly achieved on several probabilistic genome models.**

I. Introduction

DNA sequencing is the basic workhorse of modern day biology and medicine. Since the sequencing of the Human Reference Genome ten years ago, there has been an explosive advance in sequencing technology, resulting in several orders of magnitude increase in throughput and decrease in cost. Multiple "next-generation" sequencing platforms have emerged. All of them are based on the whole-genome shotgun sequencing method, which entails two steps. First, many short reads are extracted from random locations on the DNA sequence, with the length, number, and error rates of the reads depending on the particular sequencing platform. Second, the reads are assembled to reconstruct the original DNA sequence.

Assembly of the reads is a major algorithmic challenge, and over the years dozens of assembly algorithms have been proposed to solve this problem [1]. Nevertheless, the assembly problem is far from solved, and it is not clear how to compare algorithms nor where improvement might be possible. Recent work has begun to address this issue by studying the information theoretic limit of DNA assembly [2], [3], [4], [5]. These works characterize the read length, the number of reads and how low the noise level is needed to guarantee successful assembly, as a function of the repeat complexity of the genome. [2], [3] look at the case of noiseless reads and [4], [5] extend these results to noisy reads. However, all of these works assume that each read is a contiguous sub-sequence of the DNA sequence. For many sequencing platforms, each read is in fact a pair of subsequences separated by a certain distance, see Figure 1. For example, in the Illumina platform, *paired-end* reads can be obtained, with each paired-end read consists of two subsequences of length say 100 nucleotides separated by 150 nucleotides. In more laborious sample preparations, *mate-pairs* can be obtained with the paired sequences separated by up to a few thousand nucleotides. Typically, the separation between the paired subsequences is random but has small variance.

The information theoretic limit of assembly with paired reads is a much more difficult problem than that with single reads. To make progress, we make several simplifying assumptions: 1) no read noise; 2) non-random and known separation D, same for each paired read; 3) infinite number of paired reads. The last assumption is equivalent to knowing the *spectrum* of the DNA sequence, i.e. the availability of paired reads from every starting location of the DNA sequence. This assumption was very useful in solving the general problem in the case of single reads, as it allowed us to focus on the condition on the read length L for successful assembly without worrying about the number of reads first. In the paired read case, it allows us to focus on the condition on the read length and the separation.

The main contributions of this paper are:

- We show that while the problem of DNA assembly with single reads can be cast as solving 1-D jigsaw puzzles, the problem of DNA assembly with paired reads is most naturally cast as solving 2-D jigsaw puzzles.
- Using this representation, a necessary condition for successful assembly is obtained. This is a generalization of Ukkonen's condition [6] for assembly with single reads.
- Sufficient conditions for successful assembly are derived, and they are shown to be close to the necessary conditions for several probabilistic models for genomes.

There is a belief in the assembly field that paired reads are as good as having long single reads with the gap between the pair filled in, see, for example [7]. Our result provides some support to this belief.

II. Model

Let $\mathbf{s} = s_1, \ldots, s_G$ be a sequence of G symbols in $\{A, C, G, T\}$. Define $\mathbf{s}_t^\ell = s_t, \ldots, s_{t+\ell-1}$, $t = 1, \ldots, G$, where indices are modulo-G. Consider the multiset of all pairs of length-$L/2$ subsequences of $\mathbf{s}$ that are D positions apart (see Figure 1): $R(\mathbf{s}) = \left\{ r_i = (\mathbf{s}_i^{L/2}, \mathbf{s}_{i+L/2+D}^{L/2}) \colon i = 1, \ldots, G \right\}$. We will call r_i the ith (paired) read, and $R(\mathbf{s})$ the (L, D)-*spectrum of* $\mathbf{s}$. Note that for $D = 0$, $R(\mathbf{s})$ is equivalent to the

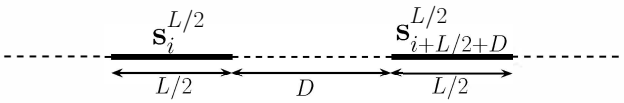

Fig. 1.

978-1-4799-5186-4/14/$31.00 ©2014 IEEE

Fig. 2. A triple repeat (left) and an interleaved repeat (right). Regions A and B are swappable, and so are C and D.

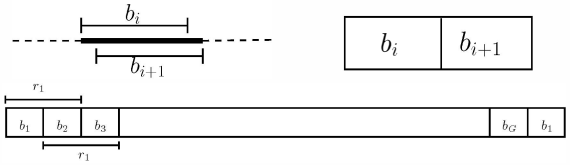

Fig. 3. Reads can be thought of as 1×2 jigsaw puzzle pieces (top), and the sequence as the solution of the jigsaw puzzle (bottom).

spectrum of single reads of length L. Consider now the set of sequences that are uniquely determined by their (L, D)-spectrums, that is, the set of $\mathbf{s}$'s for which $\mathbf{s}' \neq \mathbf{s}$ implies $R(\mathbf{s}') \neq R(\mathbf{s})$. A problem of interest is to find a simple characterization of this set.

A. Single reads: $(L, 0)$

Such a characterization is known for the $(L, 0)$-spectrum of single reads $r_i = \mathbf{s}_i^L$: a sequence is uniquely determined by its $(L, 0)$-spectrum if and only if the sequence has no interleaved or triple repeats of length at least $L - 1$. A *repeat* of length ℓ is a pair of subsequences $\mathbf{s}_i^\ell = \mathbf{s}_j^\ell$ with $i \neq j$, such that $s_{i-1} \neq s_{j-1}$ and $s_{i+\ell} \neq s_{j+\ell}$; s_i^ℓ and s_j^ℓ are the *copies* of the repeat. A triple repeat has three copies, and an interleaved repeat consists of two repeats with interleaved copies (Figure 2). The necessity of this condition (sometimes called Ukkonen's condition) can be seen as follows [6]: Write each read, slightly abusing the notation, as a pair $r_i = (b_i, b_{i+1})$, where $b_i = \mathbf{s}_i^{L-1}$. These two components can be viewed as cells of a 1×2 jigsaw puzzle piece as in Figure 3 (top). Neighboring pieces r_i and r_{i+1} overlap in b_{i+1}, and therefore can be merged as in Figure 3 (bottom). Merging all pieces in this fashion yields a solved 1-D jigsaw puzzle, and reading the symbols from left to right yields $\mathbf{s}$. Triple and interleaved repeats of cells in this solution create regions that can be *swapped* to obtain another valid solution of the puzzle. See Figure 3 (bottom) for an illustration. It was also shown in [8] that a given solution is unique if it has no swappable regions, i.e., that Ukkonen's condition is also sufficient.

B. Paired reads with arbitrary (L, D)

We wish to find an analogue to Ukkonen's condition for arbitrary (L, D)-spectrums. Following the single read case, one can write r_i, again abusing the notation, as a quadruple

$$r_i = (a_i, a_{i+1}, a_{i+L/2+D}, a_{i+L/2+D+1}), \text{ where } a_i = \mathbf{s}_i^{L/2-1}.$$

Figure 4(a) shows the four components of r_i, and Figure 4(b) shows them written in a 2×2 tile. Similar to the single read case, the right column of r_i is identical to the left column of r_{i+1}, and the bottom row of r_i is identical to the top row of $r_{i+L/2+D}$, see Figure 5(a). Therefore these reads can be merged as in Figure 5(b). Merging all neighboring reads of the spectrum in this fashion, one obtains this time a *torus*, depicted in Figure 6. To represent this torus in 2-D, we copied the last cell in each row to the first cell in the next row. In addition, the first $L/2+D+1$ cells are repeated at the end. That is, this torus is obtained by first wrapping the string $a_1 a_2 \ldots a_G a_1$ into a helix (which makes a_i and $a_{i+L/2+D}$ vertically adjacent), and then merging the two ends of this helix (which makes r_G and r_1 horizontally adjacent). The underlying sequence $\mathbf{s}$ is obtained by reading the symbols in each row of the figure in succession, from left to right. Valid solutions of this toric jigsaw puzzle by 2×2 pieces from the spectrum $R(\mathbf{s})$ yield all sequences with (L, D)-spectrum $R(\mathbf{s})$.

We now describe a necessary condition for such a jigsaw puzzle to have a unique solution. Some notation first[1]: Given a sequence $\mathbf{s}$, two cells a_i and a_j are said to be *neighbors* if they are both components of some read r_k. Therefore, each cell has 8 neighbors (Figure 5(b)). Connectedness of a set of cells is defined through this neighborhood definition, and a connected set $A \subset [1 : G]$ is *solid* if it has no 'holes'. A cell $a_i \notin A$ is on the *boundary* of A if it has a neighbor $a_j \in A$. Two solid regions are called *swappable* if their boundaries are identical (Figure 7). Note the similarity of these swappable regions to those in the 1-D case discussed above; swapping the reads that tile two such regions yields another valid solution of the puzzle. This implies

Proposition 1. *If the jigsaw puzzle representation of* $\mathbf{s}$ *has nonidentical swappable regions, then there is a sequence* $\mathbf{s}' \neq \mathbf{s}$ *with* $R(\mathbf{s}') = R(\mathbf{s})$.

Given the analogy between the definitions for 1-D and 2-D puzzles, it is tempting to conjecture that the condition above is sufficient, i.e., a solution to a 2-D jigsaw puzzle is unique if it has no swappable regions. The problem remains open.

On the other hand, the sufficiency of Ukkonen's condition in 1-D also implies a sufficient condition in 2-D. In particular, if in merging the 2×2 pieces one only requires the horizontally neighboring pieces (i.e., r_i and r_{i+1}) to be compatible, disregarding vertical neighbors, then one obtains a 1-D puzzle with 1×2 pieces, where each cell is a vertical pair $(a_i, a_{i+L/2+D})$.

[1]The results here generalize to jigsaw puzzles of arbitrary shapes once the definitions are modified to handle boundary effects.

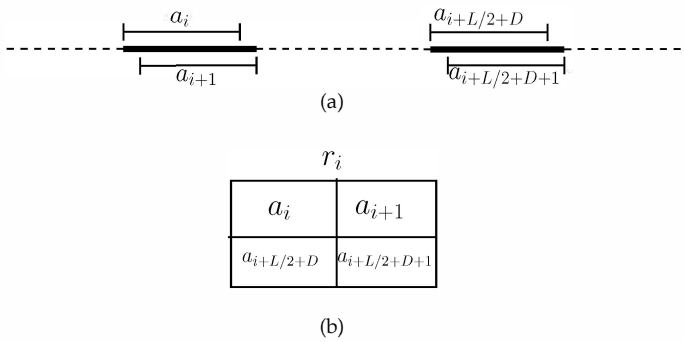

Fig. 4. Four components of the paired-read r_i (top). The same read written as a 2×2 tile (bottom).

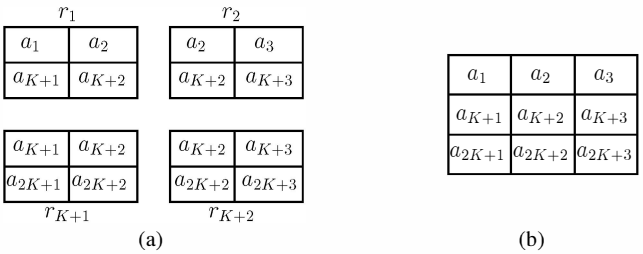

Fig. 5. (a) 2×2 representation of four neighboring reads. Here, $K = L/2 + D$. (b) Four 2×2 reads are merged to tile a 3×3 region.

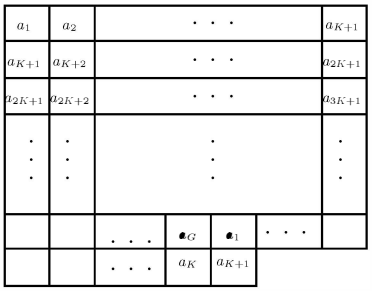

Fig. 6. Merging all paired reads of the spectrum $R(\mathbf{s})$ results in a torus. Reading the symbols from left to right yields the sequence. Here, $K = L/2+D$.

Clearly, if this puzzle has a unique solution, then it must correspond to the original sequence. This implies

Proposition 2. *A sequence* $\mathbf{s}$ *is uniquely determined by* $R(\mathbf{s})$ *if it has no interleaved or triple repeats of* $(a_i, a_{i+L/2+D})$.

We now describe another useful and relatively simple sufficient condition. Suppose that the jigsaw puzzle representation of $\mathbf{s}$ can be partitioned into regions $B, A_1, \ldots, A_m \subseteq [1:G]$ with the following properties (see Figure 7):

(i) Every A_k has at least one other copy $A_l = A_k$.
(ii) There are no repeats of a_i's apart from those implied by (i).
(iii) Each A_k's boundary lies entirely in B.

Proposition 3. *A sequence is uniquely determined by* $R(\mathbf{s})$ *if its jigsaw puzzle representation has no swappable regions and satisfies (i)–(iii).*

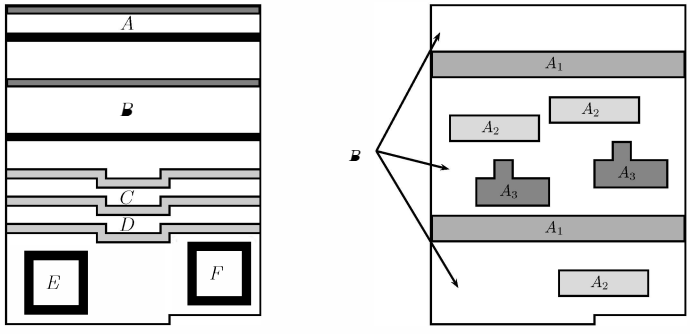

Fig. 7. (left) Boundary regions of identical shape and color consist of identical symbols. Therefore, regions A and B are swappable. So are C and D, and E and F. (right) A jigsaw puzzle that has no swappable regions and that satisfies (i)–(iii).

Remark. *Propositions 2 and 3 do not imply each other.*

Proof sketch: Conditions (i)–(iii) imply that B consists of unique cells. Therefore, each of its connected components can be assembled from the reads, including boundaries, which consist of cells in the A_k's. Similarly, it follows from (i) and (ii) that reads that belong to each A_k can be determined and each A_k can be assembled from these and attached to B. Note that the 'no swappable regions' condition implies that the A_k's are solid. This procedure yields the correct sequence if B is connected. Otherwise, it assembles a solid region for each connected component of B. It follows from the conditions that these regions are cylinder-like, as in the three regions of B including the A_k's in Figure 7. Each region is identified by its top and bottom borders, and therefore reconstructing the sequence from these regions is equivalent to solving a 1-D puzzle with 1×2 pieces, where each piece is described by a pair of borders. Since the original 2-D puzzle has no swappable regions, neither does the 1-D puzzle, and therefore the latter has a unique solution. ■

It can be seen that conditions (i)–(iii) are not necessary for unique solutions, and can be relaxed in several directions. For instance, one can replace (ii) by the recursive condition that $B, A_1, \ldots, A_m$ do not share any cell values and themselves consist of regions that satisfy (i)–(iii). Minor modifications to the argument above proves the sufficiency of this condition.

III. Bounds for probabilistic settings

We now use the observations above to derive conditions on (L, D) pairs for unique reconstruction under various probabilistic assumptions on the sequence. We first assume that the DNA is independent and identically distributed (i.i.d.) Although this is model is oversimplistic, the analysis here is a building block of the analysis of more realistic models we discuss later.

A. Independent, identically distributed (i.i.d.) sequences

Suppose that $\mathbf{s}$ is i.i.d. $\sim p$. Recall that when the sequence length G is large, the typical length of the longest interleaved repeat in $\mathbf{s}$ is roughly

$$\ell_{\text{iid}} := \bar{\ell}_{\text{iid}} \log G, \quad \text{where} \quad \bar{\ell}_{\text{iid}} = \frac{2}{H_2(p)}$$

and $H_2 = -\log \sum_i p(i)^2$ is the Renyi entropy of p [2]. This implies that assembly from the $(L, 0)$-spectrum requires

$$\bar{L} := \frac{L}{\log G} > \bar{\ell}_{\text{iid}}.$$

This argument can be generalized to arbitrary (L, D):

1) Outer bound: Consider the length of the shortest possible boundary that can have two distinct interiors. The necessary condition in Proposition 1 requires this length to be longer than ℓ_{iid}, since if a such boundary is repeated, then with high probability (w.h.p.) its two copies will have distinct swappable interiors. Observe that the shortest connected boundary of this

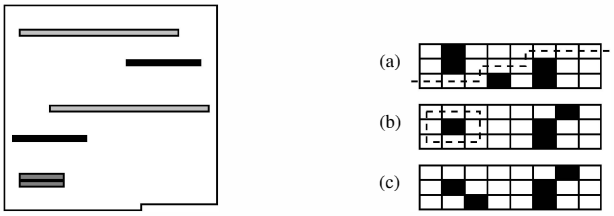

Fig. 8. (left) Three long repeats with two copies each (left). (right) Three examples of approximate repeats of length 22, with $L/2 = 2$ and $D = 5$. White cells mark the locations that are identical in the two copies, black cells mark SNP locations.

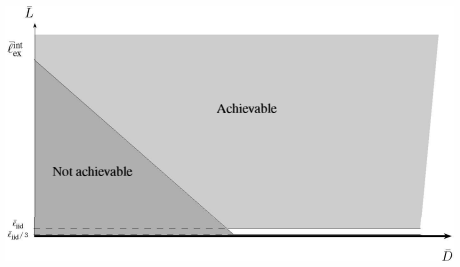

Fig. 9. Bounds on $\bar{L}$ and $\bar{D}$ for exact repeats, with $\ell_{iid} \approx 20$ and $\ell_{\text{ex}}^{\text{int}} = 3,314$ (*Lactobasillus acidophilus*). The gap is exaggerated for visibility.

type (e.g., boundary of E in Figure 7) spans 3 rows and $L/2+1$ columns, the latter due to every symbol s_i appearing in $L/2-1$ consecutive cells. There are $3L - 4 \approx 3L$ symbols in such a boundary. Note also that interleaved, length-$(L/2+D)$ repeats of consecutive cells, i.e., $L + D - 2 \approx L + D$ symbols, also yield swappable regions with unconnected boundaries (e.g., A and B in Figure 7). It then follows that w.h.p., there will be no swappable regions if

$$\begin{aligned} \bar{L} &> \bar{\ell}_{\text{iid}}/3 \\ \bar{L} + \bar{D} &> \bar{\ell}_{\text{iid}}. \end{aligned} \tag{1}$$

2) Inner bound: Consider the sufficient condition in Proposition 2. Since each $(a_i, a_{i+L/2+D})$ pair contains $L - 2 \approx L$ symbols, it suffices to have

$$\bar{L} = \frac{L}{\log G} > \bar{\ell}_{\text{iid}}$$

for each such pair to be unique and thus the condition to be satisfied.

B. I.I.D. background with long exact repeats

As observed in [2], the symbol statistics of most DNA sequences are rarely typical of i.i.d. distributions. In particular, an i.i.d. $\sim p$ distribution often explains a large part of a given DNA, although there are typically a number of repeats that are much longer than the i.i.d. model would predict. That is, the length of the longest repeat, $\ell_{\text{ex}}^{\text{int}}$, is often much larger than ℓ_{iid}. For example, in the *Lactobasillus acidophilus* genome of length $G = 2,078,001$, we have $\ell_{\text{iid}} \approx 20$, and $\ell_{\text{ex}}^{\text{int}} = 3,314$. Perhaps the simplest way to guarantee such long repeats in a probabilistic setting is to generate them explicitly (see Figure 8):

- Given a distribution p, fix $\bar{\ell}_n \geq \ldots \geq \bar{\ell}_2 \geq \bar{\ell}_1 := \bar{\ell}_{\text{iid}}$, and define $\ell_i = \bar{\ell}_i \log G$. Pick, uniformly at random, cn non-overlapping sections $\mathbf{s}_{n_{ij}}^{\ell_i}$, $i = 1, \ldots, n$, $j = 1, \ldots, c$ of $\mathbf{s}$ and set
$$\mathbf{s}_{n_{ij}}^{\ell_i} = \mathbf{x}_i,$$
where $\mathbf{x}_i$'s are independent, i.i.d. $\sim p$ sequences of length ℓ_i. Note that each $\mathbf{x}_i$ has c copies in $\mathbf{s}$.
- Pick the remaining s_i's i.i.d. $\sim p$.

1) Outer bound: Note first that the total length of the long repeats $\mathbf{s}_{n_{ij}}^{\ell_i}$ vanish as a fraction of G, i.e., almost all symbols in $\mathbf{s}$ are i.i.d. It thus follows that the conditions.

$$\begin{aligned} \bar{L} &> \bar{\ell}_{\text{iid}}/3 \\ \bar{L} + \bar{D} &> \bar{\ell}_{\text{iid}}. \end{aligned}$$

for the i.i.d. setting above are also necessary here. Let $\bar{\ell}_{\text{ex}}^{\text{int}}$ denote the length of the longest triple or interleaved repeat in $\mathbf{s}$. We have w.h.p. $\bar{\ell}_{ex}^{\text{int}} = \bar{\ell}_i$, for some i. Observe again that if $\bar{\ell}_{\text{ex}}^{\text{int}} \geq \bar{L} + \bar{D}$, then there will be swappable regions, similar to A and B in Figure 7. Hence,

$$\bar{L} + \bar{D} > \bar{\ell}_{\text{ex}}^{\text{int}}$$

is also necessary. Note that this renders the bound $\bar{L}+\bar{D} > \bar{\ell}_{\text{iid}}$ above redundant since $\bar{\ell}_{\text{ex}}^{\text{int}} \geq \bar{\ell}_{\text{iid}}$.

2) Inner bound: Consider again the sufficient condition in Proposition 2. Note that for all i, the $L-2 \approx L$ symbols in the vertical pair $(a_i, a_{i+L/2+D})$ are independent w.h.p. (i.e., unless two copies of a repeat are vertically adjacent as in the bottom repeat in Figure 8. The probability of this event is small since the repeat locations are chosen uniformly at random.) Also, if $\bar{L} + \bar{D} > \bar{\ell}_{\text{ex}}^{\text{int}}$, then all repeats will span less than one row. Therefore, if we have in addition $\bar{L} > \bar{\ell}_{\text{iid}}$, then all vertical $(a_i, a_{i+L/2+D})$ pairs will be unique w.h.p. That is, the bounds

$$\begin{aligned} \bar{L} + \bar{D} &> \bar{\ell}_{\text{ex}}^{\text{int}} \\ \bar{L} &> \bar{\ell}_{\text{iid}}, \end{aligned} \tag{2}$$

guarantee that the sufficient condition in Proposition 2 is satisfied. The inner and the outer bounds, are shown in Figure 9. Note that even if the D symbols between the paired reads were known for every read, Ukkonen's condition would imply that $\bar{L} + \bar{D} > \bar{\ell}_{\text{ex}}^{\text{int}}$ is necessary for unique solutions. At the other extreme, treating both reads of a pair as single reads requires $\bar{L}/2 > \bar{\ell}_{\text{ex}}^{\text{int}}$. Since $\ell_{\text{iid}} \ll \ell_{\text{ex}}^{\text{int}}$, bounds in (2) imply that paired short reads are almost as useful as much longer single reads.

C. I.I.D. background with long approximate repeats

Often, the long repeats in a DNA are not exact, i.e., two copies of a repeat differ in a small number of locations; these differences are sometimes called single-nucleotide polymorphisms (SNP). For example, in *Salmonella enterica*, the length of the longest interleaved *approximate* repeat is $\ell_{\text{app}}^{\text{int}} \approx 600$.

Within this repeat, there are 24 SNPs, and the longest inter-SNP distance is, 163 which yields $\ell_{\text{ex}}^{\text{int}} = 163$. We now discuss the effect of SNPs on the long exact repeats in the previous setting.

1) Outer bound: It is easily seen that the necessary conditions

$$\bar{L} > \bar{\ell}_{\text{iid}}/3 \tag{3}$$

$$\bar{L} + \bar{D} > \bar{\ell}_{\text{ex}}^{\text{int}} \tag{4}$$

above for exact repeats are also necessary here.

2) Inner bound: Consider first treating approximate repeats as exact repeats. Then, as in the previous section, having

$$\bar{L} > \bar{\ell}_{\text{iid}} \tag{5}$$

$$\bar{L} + \bar{D} > \bar{\ell}_{\text{app}}^{\text{int}} \tag{6}$$

is sufficient for assembly, where $\bar{\ell}_{\text{app}}^{\text{int}}$ is the length of the longest approximate interleaved/triple repeat.

Note that the gap between (4) and (6) can be significant since, as mentioned above, $\ell_{\text{app}}^{\text{int}}$ is often much larger than $\bar{\ell}_{\text{ex}}^{\text{int}}$. The gap between (3) and (5) is less significant since ℓ_{iid} is small, see Figure 10 (left). It is therefore of interest to determine whether the region $\bar{\ell}_{\text{ex}}^{\text{int}} < \bar{L} + \bar{D} < \bar{\ell}_{\text{app}}^{\text{int}}$ contains any achievable $(\bar{L}, \bar{D})$ pairs. For this purpose, observe first that the sufficient condition in Proposition 2 is not useful when $\bar{\ell}_{\text{app}}^{\text{int}} > \bar{L} + \bar{D}$. We therefore turn to Proposition 3. Suppose that $\bar{L}/2 > \bar{\ell}_{\text{iid}}$, and thus all i.i.d. cells are unique w.h.p. As before, long repeats will be separated by at least one row of i.i.d. cells, and thus the sequence will satisfy conditions (i)–(iii), where the set of i.i.d. cells correspond to B, and the repeat regions correspond to the A_k's. It then remains to determine whether the repeats themselves lead to swappable regions. This is best done through an example: Figure 8 depicts three different approximate repeats, each spanning three rows. Black cells mark SNPs; white cells are identical in each copy of the repeat. Consider Figure 8(a). Suppose that this copy belongs to one of the two interleaved approximate repeats, and for simplicity, that both repeats have the same length and identical SNP locations. Since both copies of both repeats are identical on the dashed path, these repeats create two swappable regions (similar to A and B in Figure 7). Similarly, in Figure 8(b), the interior of the dashed path in the two copies are swappable since they have identical boundaries (similar to E and F in Figure 7). On the other hand, SNPs in Figure 8(c) do not yield swappable regions. For unique assembly, then, approximate interleaved repeats of length $L+D$ must be such that

(I) at least one of the two interleaved repeats is 'bridged' by a set of connected SNPs, such as the three on the right sides of Figures 8(b) and (c),

(II) there are no isolated SNPs in either repeat, such as the one on the left of Figure 8(b).

Note that the structure of SNP locations, i.e., whether the above conditions hold, is sensitive to the values of L and D. See Figure 10 for an example, which plots the feasible and unfeasible regions for a particular approximate repeat in *Salmonella enterica.*

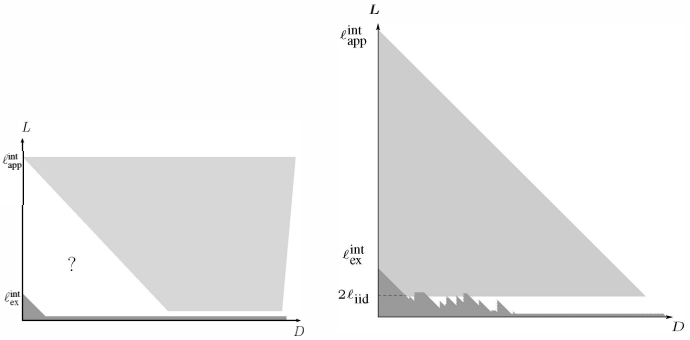

Fig. 10. Achievable (light) and unachievable (dark) regions for an approximate interleaved repeat in *Salmonella enterica* with $\ell_{\text{ex}}^{\text{int}} = 94$, and $\ell_{\text{app}}^{\text{int}} = 540$, assuming no other long approximate repeats, and $\ell_{\text{iid}} = 20$. Only the region $L + D < \ell_{\text{app}}^{\text{int}}$ is shown.

In the present model, the necessary conditions above, in addition to similar ones for triple repeats, and $L/2 > \bar{\ell}_{\text{iid}}$, are sufficient. In this example, a large fraction of the region $\ell_{\text{ex}}^{\text{int}} < L + D < \ell_{\text{app}}^{\text{int}}$ is achievable.

Acknowledgement

This work is supported by the Swiss National Science Foundation under grant PBELP2_137726 and by the NSF Center for Science of Information under grant agreement CCF-0939370.

References

[1] Wikipedia, *Sequence assembly – Wikipedia, the free encyclopedia,* 2012, [Online] http://en.wikipedia.org/wiki/Sequence_assembly.

[2] S. A. Motahari, G. Bresler, D. Tse, "Information theory of DNA shotgun sequencing," *IEEE Trans. Inform. Theory,* vol. 59, no. 10, pp. 6273–6289, 2013.

[3] G. Bresler, M. Bresler, D. Tse, "Optimal assembly for high throughput shotgun sequenceing," *BMC Bioinformatics,* vol. 14, 2013.

[4] S. A. Motahari, K. Ramchandran, D. Tse, N. Ma, "Optimal DNA shotgun sequencing: noisy reads are as good as noiseless reads," *Proc. Int. Symp. Inform. Theory,* 2013.

[5] K.-K. Lam, A. Khalak, D. Tse, "'Near-optimal assembly for shotgun sequencing with noisy reads, to appear in BMC Bioinformatics, 2014.

[6] E. Ukkonen, "Approximate string matching with q-grams and maximal matches," *Theoretical Computer Science,* vol. 92, no. 1, pp. 191-211, 1992.

[7] M. J. Chaisson, D. Brinza, P. A. Pevzner, "De novo fragment assembly with short mate-paired reads: does the read length matter?" *Genome Research*, vol. 19, no. 2, pp.336–346, 2009.

[8] P. A. Pevzner, "DNA physical mapping and alternating Eulerian cycles in colored graphs," *Algorithmica,* vol. 13, no. 1/2, pp. 77–105, 1995.