Fast scaffolding with small independent mixed integer programs

Salmela, Leena

2011


http://hdl.handle.net/10138/29432
https://doi.org/10.1093/bioinformatics/btr562

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.
Although there are good methods to assemble reads produced by the second-generation sequencing machines, because the data may be incomplete and the organism may have several chromosomes. Instead each contig is assigned to a scaffold and given an orientation and position within the scaffold.

Kececioglu and Myers (1995) have shown that even determining the orientation of the contigs is NP-hard. Therefore, all practical methods to solve the scaffolding problems use heuristics and achieve only an approximate solution.

Many assemblers like Velvet (Zerbino and Birney, 2008), Allpaths (Butler et al., 2008) and SOAPdenovo (Li et al., 2010) contain a scaffolding module. Some stand-alone scaffolders have also been developed. Bambus (Pop et al., 2004) is designed for Sanger data, and SOPRA (Dayarian et al., 2010) and SSPACE (Boetzer et al., 2011) are developed for second-generation sequencing data. Bambus and SSPACE are based on a greedy method, whereas SOPRA relies on statistical optimization and partitioning the scaffolding problem. Our new approach uses a partitioning scheme similar to SOPRA but unlike SOPRA we restrict the size of the partitions and thus we can solve the subproblems exactly with mixed integer programming.

The rest of the article is outlined as follows. In Section 2, we present a theoretical framework for scaffolding based on mixed integer programming and partitioning the scaffolding problem. Section 3 presents a practical implementation of this framework. We compare the new scaffolder to SOPRA and SSPACE in Section 4. Finally, we conclude in Section 5.

1 INTRODUCTION

Present high-throughput sequencing machines can produce hundreds of millions of short reads in a single run. Initially, this technology was targeted at resequencing applications but is nowadays also deployed in de novo sequencing projects. Apart from single reads, the sequencing machines can also produce mate pairs, i.e. pairs of reads whose approximate distance in the target genome is known. Although there are good methods to assemble reads produced by the older Sanger technology, new approaches are needed to deal with data from the second-generation sequencing machines, because the characteristics of the data like read length and sequencing errors are different (Pop, 2009).

In a de novo sequencing project, the read data is typically first quality trimmed and filtered to ensure high-quality data for subsequent processing. A method to correct sequencing errors may also be used. The next step is to assemble the reads into contigs, which are gapless sequences of nucleotides. In the last step, the contigs are ordered into scaffolds based on mate pair reads (Pop, 2009).

Here we consider the scaffolding problem. The input to the scaffolding problem is the contigs produced by a contig assembler, the mappings of the mate pair reads to the contigs and the insert sizes of the mate pair libraries. The objective is to find a linear ordering of the contigs, which maximizes the number of mate pairs whose pairwise distance equals the insert size. In practise, a linear ordering of all the contigs is not achieved, because the data may be incomplete and the organism may have several chromosomes. Instead each contig is assigned to a scaffold and given an orientation and position within the scaffold.

Kececioglu and Myers (1995) have shown that even determining the orientation of the contigs is NP-hard. Therefore, all practical methods to solve the scaffolding problem use heuristics and achieve only an approximate solution.

Many assemblers like Velvet (Zerbino and Birney, 2008), Allpaths (Butler et al., 2008) and SOAPdenovo (Li et al., 2010) contain a scaffolding module. Some stand-alone scaffolders have also been developed. Bambus (Pop et al., 2004) is designed for Sanger data, and SOPRA (Dayarian et al., 2010) and SSPACE (Boetzer et al., 2011) are developed for second-generation sequencing data. Bambus and SSPACE are based on a greedy method, whereas SOPRA relies on statistical optimization and partitioning the scaffolding problem. Our new approach uses a partitioning scheme similar to SOPRA but unlike SOPRA we restrict the size of the partitions and thus we can solve the subproblems exactly with mixed integer programming.

The rest of the article is outlined as follows. In Section 2, we present a theoretical framework for scaffolding based on mixed integer programming and partitioning the scaffolding problem. Section 3 presents a practical implementation of this framework. We compare the new scaffolder to SOPRA and SSPACE in Section 4. Finally, we conclude in Section 5.

2 ALGORITHM

The input to a scaffolder is a set of contigs produced by an assembler and a set of mate pairs. First, we map the mate pairs to the contigs. This yields links between contigs that we represent as a graph. The graph is then partitioned into subproblems, which we solve with mixed integer programming. Finally, we combine the solutions of the subproblems to solve the whole scaffolding problem. Each of these tasks is discussed in more detail below. In Section 3, we address practical issues in implementing our approach.

2.1 Scaffolding graph

The input to the scaffolding problem can be represented as a bidirected graph called the scaffolding graph. Each contig is represented by a node in the graph. There is an edge between two contigs if there are mate pairs linking them. The edge is directed at both endpoints indicating the orientation of the contigs with regard to each other. Figure 1 shows the four possible ways of linking two contigs. Here the mate pair ends are oriented as in SOLiD data, where both reads of the pair are from the same strand of the genome.
Several mate pairs can induce the same edge in the scaffolding graph. The total number of these mate pairs is the support $s_k$ of the edge $k$. These mate pairs can be used to estimate the distance $d_k$ between the two contigs.

The scaffolding problem can now be formulated as removing a set of edges with minimum combined support from the scaffolding graph such that each node can be assigned a position and orientation that satisfy the constraints imposed by the remaining edges. Each scaffold can then be constructed in two equivalent ways: the other way can be obtained by reversing the whole scaffold.

2.2 Mixed integer programming formulation

In this section, we will develop a mixed integer programming (MIP) formulation for the scaffolding problem. As a convention, we will use italic type to denote variables and roman type to denote constants for those formulation for the scaffolding problem. As a convention, we will use italic type to denote variables and roman type to denote constants for those.

Fig. 1. The four possible ways of linking two contigs with a mate pair and the bidirected edge representing each case.

Fig. 2. The two possible placements of contigs $i$ and $j$ when an edge with orientation corresponding to the top left case in Figure 1 is used in a scaffolding. The top figure shows the case when both contigs are forward oriented and the bottom one shows the case when they are reverse oriented.

From the top case of Figure 2, we see that the following should hold when both contigs are forward oriented:

$$s_i + s_j + d_k = s_{ij}$$

We get the following relaxed constraints:

$$s_i + a_k + d_k - C(1 - s_k) - C(1 - a_k) \leq s_{ij}$$

$$s_i + a_k + d_k + C(1 - s_k) + C(1 - a_k) \geq s_{ij}$$

where $C$ is a large enough constant to relax the constraints if either edge is not included in the solution or the two contigs are reverse oriented. A good choice for $C$ is 2L as $s_k \leq L$ and $s_j \leq L$. Similarly, we get the relaxed constraints for the case where both contigs are reverse oriented (see bottom of Fig. 2):

$$x_i + d_k + a_k - C(1 - s_k) - Cn \leq s_{ij}$$

$$x_i + d_k + a_k + C(1 - s_k) + Cn \geq s_{ij}$$

The constraints for edges corresponding to the other three cases shown in Figure 1 are derived similarly. For each edge, we thus get six constraints corresponding to Equations (1–6). In total, we get 6w constraints where $w$ is the number of edges in the scaffolding graph.

When defining the constraints imposed by edge $k$, we allowed the distance $d_k$ to vary by $C(1 - s_k)$. Our objective is to minimize the amount by which the solution to the MIP problem assigns values to the variables $x_i$ and $x_j$.

Additional, there is one real-valued variable for each edge in the scaffolding graph.

Fig. 2 shows the two possible placements of the contigs $i$ and $j$ with orientation corresponding to the top left case in Figure 1 is used in a scaffolding. The top figure shows the case when both contigs are forward oriented and the bottom one shows the case when they are reverse oriented.

2.3 Partitioning the scaffolding problem

Dayarian et al. (2010) noticed that the scaffolding problem can be partitioned into subproblems that can be solved independently. If the removal of a contig divides the scaffolding graph into two components, then the scaffolding can be solved independently for these two components both of which also
To estimate the distance $d$

We first use a read mapper like the readaligner tool by Mäkinen

The output of MIP Scaffolder is the scaffolds in a FASTA file.

contains $r$

be found by sliding a window of length $w$

and link the contig to the same direction, then at least $p$

parameters of our method. Figure 3 shows examples of ($w$, 0.5)-consistent and inconsistent mate pair mappings. The arrows denote contigs and the short lines mate pairs. The mate pair ends are connected with arcs. The mate pairs with solid line arcs connecting them are consistent and the mate pairs with dashed line arcs connecting them are not consistent.

$\text{MIP Scaffolder}$

3 IMPLEMENTATION

3.1 General

MIP Scaffolder is implemented as a set of C++ programs. We use Lemon graph library (http://lemon.cs.elte.hu) to implement the scaffolding graph and lp_solve (http://lpsolve.sourceforge.net/) to solve the MIP problems.

As input MIP Scaffolder takes the contigs as FASTA files and the alignments of the mate pairs to the contigs as SAM files. MIP Scaffolder can use several mate pair libraries with different insert sizes simultaneously. The output of MIP Scaffolder is the scaffolds in a FASTA file.

3.2 Filtering mate-pair mappings

We first use a read mapper like the readaligner tool by Mäkinen et al. (2010) to map the mate pairs to the contigs. For further processing, we use only those mate pairs whose both ends map to a unique position in the contig collection.

To improve the quality of the mate pair mappings, we filter the mappings so that we only keep mappings that are consistent with other mate pairs mapping to approximately the same location.

We call a mate pair mapping ($w$, $p$)-consistent if for both ends of the mate pair the following holds: if a mate pair end maps to a contig, then there exists a window of length $w$ on this contig such that if in total $r$ mate pair ends map into the same window and link the contig to the same direction, then at least $p \cdot r$ of all these mate pairs link the same pair of contigs as the original mate pair. We only keep the mate pairs mappings that are ($w$, $p$)-consistent for some $w$ and $p$ which are parameters of our method. Figure 3 shows examples of ($w$, 0.5)-consistent and inconsistent mate pair mappings.

Consistent mate pair mappings can be found by sliding a window of length $w$ on each contig. If the window contains $r$ mappings and $p \cdot r$ of these link the same pair of contigs, then the $p \cdot r$ mate pair ends are marked consistent. We then scan all the mate pair mappings and keep those whose both ends are marked consistent.

Filtering out mate pairs that are not consistent has two benefits. First, it means that mate pairs mapping to repeat regions are not used in the scaffolding process even if the repeat region is present only once in the contig collection.

If all repeat regions would be present several times, keeping only unique mappings should suffice. Second, the insert length distributions tend to have long tails. Mate pairs whose insert length deviates significantly from the mean may be filtered out allowing us to estimate the distance between contigs more accurately.

3.3 The scaffold graph

To estimate the distance $d_	ext{min}$ associated with an edge $k$ linking two contigs, we use edge bundling (Pop et al., 2004) combined with a maximum likelihood method (Gnecco et al., 2011; Simpson et al., 2009).

For each mate pair library, we estimate the mean insert size $\text{Emean}$ and the insert size range $[\text{Emin}, \text{Emax}]$. Mate pairs whose insert length deviates significantly from their high degree in the scaffolding graph or much higher than expected coverage and remove them from the scaffolding graph. By default, contigs with coverage >2.5 times the average coverage of contigs or degree >50 are classified as repeat contigs. Both thresholds can be adjusted by the user.

4 RESULTS AND DISCUSSION

4.1 Mate pair libraries and contigs

The read sets used to test our approach are summarized in Table 1. The first two read sets are SOLID reads from the *Escherichia coli* DH110B strain and the *Caenorhabditis Elegans* genome. The *E.coli* read set is produced by Applied Biosystems, and we used the filtered subset available at http://hts.rutgers.edu/. This set is quality filtered and trimmed to 35 bp. We used the mean filter distributed with SOPRA to filter the *C.elegans* read set. Each read has an average quality at least 19 and no indeterminate colors. The third read set is an Illumina paired end library for *Pseudomonas syringae*.
L. Salmela et al.

Table 1. The test datasets

<table>
<thead>
<tr>
<th>Reference organism</th>
<th>E.coli</th>
<th>C.elegans</th>
<th>P.syringae</th>
<th>H.sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Escherichia coli</td>
<td>Caenorhabditis elegans</td>
<td>Pseudomonas syringae</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Genome size (Mbp)</td>
<td>4.6</td>
<td>100.3</td>
<td>6.1</td>
<td>3080</td>
</tr>
<tr>
<td>Accession number (SRA/ERA)</td>
<td>NA</td>
<td>SRR033650</td>
<td>ERR005143</td>
<td>SRR067771</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SRR033650</td>
<td>ERR005143</td>
<td>SRR067773</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SRR067777</td>
<td>SRR067779</td>
<td>SRR067781</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SRR067784</td>
<td>SRR067778</td>
<td>SRR067776</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SRR067776</td>
<td>SRR067786</td>
<td></td>
</tr>
<tr>
<td>Coverage</td>
<td>110×</td>
<td>10×</td>
<td>42×</td>
<td>7×</td>
</tr>
<tr>
<td>Spanning coverage</td>
<td>1930×</td>
<td>427×</td>
<td>236×</td>
<td>223×</td>
</tr>
<tr>
<td>Insert size</td>
<td>1200</td>
<td>1785</td>
<td>400</td>
<td>2500</td>
</tr>
<tr>
<td>Sequencing platform</td>
<td>SOLID</td>
<td>SOLID</td>
<td>Illumina</td>
<td>Illumina</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The statistics are given for the sets after trimming and filtering. The spanning coverage is the total coverage of the read pairs and the bases in between.

and the fourth read set consists of two Illumina mate pair libraries for the human genome. The human Illumina mate pair libraries were first trimmed to 36 bp and then we filtered out all reads where >30% of the bases had quality score ≤10. The statistics in Table 1 are for the filtered sets.

We ran the initial contig assembly with Velvet (Zerbino and Birney, 2008). The reads in the C.elegans dataset were too short and the coverage too low to assemble the reads into contigs. We were also not able to assemble contigs for the complete human genome because of memory constraints. Instead we used the wgsim tool distributed with SAMtools (Li et al., 2009) to generate synthetic reads of length 100 from the reference genome and built the contigs based on these reads. For the human genome, the reads were generated and assembled one chromosome at a time to keep the memory requirements on a reasonable level. The coverage of these synthetic reads was 30× for C.elegans. For the human genome, we generated paired reads with insert size 500 and 20× coverage. These reads were not made available for scaffolding.

When running Velvet, we kept contigs >150 bp. We used the following other parameters:

- E.coli: hash length 19 and coverage cutoff 6.
- C.elegans: hash length 61 and coverage cutoff 6.
- P.syringae: hash length 21 and coverage cutoff 6.
- H.sapiens: hash length 19, coverage cutoff 6, and expected coverage 20, no scaffolding.

The E.coli dataset was in color space and we used the contig translation of SOPRA to get contigs in base space. For the other datasets, the reads used to build the contigs were in base space and so we could directly use the contigs produced by Velvet. The statistics of the produced contigs are included in Tables 3–6.

4.2 Evaluation of scaffolds

The produced scaffolds were compared for the length of the scaffolds and for their correctness. For length comparison, we used the N50 measure as well as the total length of the scaffolds.

We evaluated the correctness of the scaffolds as follows. Using swift by Rasmussen et al. (2006), we searched for maximal approximate local matches between the scaffold and the genome sequence. With E.coli and P.syringae, we used parameters 11 for the seed length, 0.05 for the maximum error level and 30 for the minimum length of a maximal match. With C.elegans, we used the same seed length, 0.02 for the maximum error level and 35 for the minimum length of a maximal match. With the human scaffolds, we used 0.01 for the maximum error level, and 100 for the minimal length of a maximal match. This produced a set of tuples each representing a match between a substring range of scaffold and substring range of genome. To obtain an alignment for the whole scaffold inside the genome, we implemented the colinear chaining algorithm (Aboelhoda, 2007) that finds a subset of the tuples that retains the linear order in both source (scaffold) and target (genome), and maximizes the overall coverage of the source (scaffold). Additionally, we restricted the length of gaps in the colinear chaining algorithm to 2000 bp on the E.coli and P.syringae datasets and to 5000 bp on the C.elegans and human datasets. This way we could obtain a reliable alignment without resorting into laborious sequence-level dynamic programming.

Three measures were computed based on the colinear chaining. Genome coverage denotes the proportion of positions inside the genome taking part into the alignment. Scaffold coverage denotes the proportion of positions inside scaffolds taking part into the alignment. We also extracted those parts of the scaffolds that took part into the alignment and repeated the colinear chaining computation recursively for the remaining part until no parts having a maximal local approximate match remained. Finally, we computed the N50 length statistic for all extracted parts. We call this N50 value the normalized N50 statistic. Note that the extracted parts of the scaffolds do not contain gaps and so the normalized N50 statistic should be compared to the N50 value computed on the complete scaffolds excluding gaps.

The validity of scaffolds was additionally measured similarly to Gneffe et al. (2011). Pairs of sequences separated by a given distance were extracted randomly from the scaffolds. These pairs were then
mapped to the reference and we measured the proportion of pairs whose orientation matched and distance was within 10% of the given distance. The validity of contigs was not measured in this fashion as they were not long enough.

4.3 Parameters and runtime of the scaffolders

For the E.coli, C.elegans and P.syringae datasets, we used readaligner (Mäkinen et al., 2010) to map the mate pairs to contigs and for the human dataset we used SOAP2 (Li et al., 2009). For the E.coli and P.syringae datasets, the results for SOPRA are for the version which is integrated with Velvet and thus directly uses the placement of reads as designated by Velvet. We allowed at most two edit operations (mismatches and indels) when running readaligner and at most three mismatches when running SOAP2. We were able to map 91% of the E.coli reads, 87% of the C.elegans reads and 87% of the P.syringae reads. For the human dataset, SOAP2 reported a unique match for 70% of the reads.

We tried several parameter combinations for MIP Scaffold. Figure 4 shows how the N50 value and the genome coverage of the E.coli dataset evolve as we vary $w$ and $p$ when determining $(w,p)$-consistent mappings. The longest scaffolds are produced with fairly loose filtering but the loosest filtering also gives less accurate scaffolds. Figure 5 shows how the N50 value and the genome coverage evolve when the maximum size of the biconnected component is varied. Here, we notice that the maximum size of the component does not affect the scaffold lengths or correctness much. For the analysis of the scaffolds, we used $(10,0.7)$-consistent read mappings and maximum component size 100.

We similarly experimented with different parameters on the other datasets and chose those parameters that produced longest scaffolds. For the C.elegans dataset, we used $(30,0.75)$-consistent read mapping and maximum component size 50, for the P.syringae dataset $(30,0.5)$-consistent read mappings and maximum component size 50 and for the human dataset $(50,0.75)$-consistent read mappings and maximum component size 50.

We tried several parameters for running SOPRA and show the results for those parameters that gave the longest scaffolds measured by the N50 value. The E.coli and the P.syringae scaffolds were produced with the version of SOPRA that is integrated with Velvet, and the C.elegans and human scaffolds were produced using pregenerated contigs. For the E.coli dataset, we got best results with support threshold $(w,6)$. The original paper reports result with support threshold 5 but our scaffolds are slightly longer. This might be due to the stochastic nature of the algorithm which means that even different runs with the same parameters may produce different results. For the C.elegans and P.syringae datasets, we got longest scaffolds with support thresholds 20 and 4, respectively. We could not run SOPRA on the human dataset because it ran out of memory.

SSPACE does not support mate pair libraries from the SOLiD technology and so we ran it only on the P.syringae and human datasets. SSPACE was run with default parameters.

The scaffolders were run on a server with 16 cores operating at 2.93 GHz and 128 GB of memory. Table 2 shows the runtimes of the different scaffolders. We see that MIP Scaffold is the fastest of the scaffolders except on the human dataset. On the human dataset, MIP Scaffold spent half of its time preprocessing the read mappings. As further work, we plan to improve the efficiency of this phase.
L.Salmela et al.

Table 3. The length and validation statistics of scaffolds produced from the \textit{E.coli} data set by the scaffolders

<table>
<thead>
<tr>
<th>Scaffolder</th>
<th>Number of scaffolds</th>
<th>N50 (bp)</th>
<th>Total length (bp)</th>
<th>Genome coverage</th>
<th>Scaffold coverage</th>
<th>Normalized N50 (bp)</th>
<th>Validity at 10kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>4341</td>
<td>1499</td>
<td>4,362,035</td>
<td>0.922</td>
<td>0.998</td>
<td>1496</td>
<td>–</td>
</tr>
<tr>
<td>SOPRA</td>
<td>168</td>
<td>203,090</td>
<td>4,840,884</td>
<td>0.912</td>
<td>0.988</td>
<td>185,227</td>
<td>0.910</td>
</tr>
<tr>
<td>MIP Scaffolder</td>
<td>207</td>
<td>182,666</td>
<td>4,558,212</td>
<td>0.889</td>
<td>0.968</td>
<td>170,796</td>
<td>0.972</td>
</tr>
</tbody>
</table>

When computing colinear chaining to obtain genome coverage, scaffold coverage and normalized N50 values, gaps were restricted to at most 2000 bp.

Table 4. The length and validation statistics of scaffolds produced from the \textit{C.elegans} dataset by the scaffolders

<table>
<thead>
<tr>
<th>Scaffolder</th>
<th>Number of scaffolds</th>
<th>N50 (bp)</th>
<th>Total length (bp)</th>
<th>Genome coverage</th>
<th>Scaffold coverage</th>
<th>Normalized N50 (bp)</th>
<th>Validity at 10kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>31,419</td>
<td>14,717</td>
<td>96,249,755</td>
<td>0.943</td>
<td>1.000</td>
<td>14,717</td>
<td>–</td>
</tr>
<tr>
<td>SOPRA</td>
<td>17,951</td>
<td>132,547</td>
<td>98,595,934</td>
<td>0.945</td>
<td>0.999</td>
<td>130,346</td>
<td>0.990</td>
</tr>
<tr>
<td>MIP Scaffolder</td>
<td>10,721</td>
<td>10,721</td>
<td>98,273,945</td>
<td>0.933</td>
<td>0.986</td>
<td>183,891</td>
<td>0.973</td>
</tr>
</tbody>
</table>

When computing colinear chaining to obtain genome coverage, scaffold coverage and normalized N50 values, gaps were restricted to at most 5000 bp.

Table 5. The length and validation statistics of scaffolds produced from the \textit{P.syringae} dataset by the scaffolders

<table>
<thead>
<tr>
<th>Scaffolder</th>
<th>Number of scaffolds</th>
<th>N50 (bp)</th>
<th>Total length (bp)</th>
<th>Genome coverage</th>
<th>Scaffold coverage</th>
<th>Normalized N50 (bp)</th>
<th>Validity at 10kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>2251</td>
<td>6972</td>
<td>5,925,601</td>
<td>0.964</td>
<td>0.981</td>
<td>6982</td>
<td>–</td>
</tr>
<tr>
<td>SOPRA</td>
<td>568</td>
<td>75,224</td>
<td>5,987,540</td>
<td>0.958</td>
<td>0.989</td>
<td>72,714</td>
<td>0.998</td>
</tr>
<tr>
<td>MIP Scaffolder</td>
<td>103,598</td>
<td>103,352</td>
<td>5,990,318</td>
<td>0.918</td>
<td>0.949</td>
<td>84,779</td>
<td>0.983</td>
</tr>
</tbody>
</table>

When computing colinear chaining to obtain genome coverage, scaffold coverage and normalized N50 values, gaps were restricted to at most 2000 bp.

Table 6. The length and validation statistics of scaffolds produced from the \textit{H.sapiens} dataset by the scaffolders

<table>
<thead>
<tr>
<th>Scaffolder</th>
<th>Number of scaffolds</th>
<th>N50 (bp)</th>
<th>Total length (bp)</th>
<th>Genome coverage</th>
<th>Scaffold coverage</th>
<th>Normalized N50 (bp)</th>
<th>Validity at 50kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>349,514</td>
<td>18,994</td>
<td>2,785,310,700</td>
<td>0.863</td>
<td>0.959</td>
<td>18,185</td>
<td>–</td>
</tr>
<tr>
<td>SOPRA</td>
<td>97,525</td>
<td>348,941</td>
<td>2,834,193,333</td>
<td>0.640</td>
<td>0.719</td>
<td>179,418</td>
<td>0.869</td>
</tr>
<tr>
<td>MIP Scaffolder</td>
<td>83,909</td>
<td>328,665</td>
<td>2,823,814,726</td>
<td>0.684</td>
<td>0.769</td>
<td>190,008</td>
<td>0.870</td>
</tr>
</tbody>
</table>

When computing colinear chaining to obtain genome coverage, scaffold coverage and normalized N50 values, gaps were restricted to at most 5000 bp.

4.4 Comparison of scaffolds

Tables 3–6 show the length and validation statistics of the scaffolds produced by SOPRA, SSPACE and MIP Scaffolder for the various datasets. We report the scaffold lengths both including gaps and without gaps. The length statistics without gaps are more important, because the statistics with gaps can vary if the insert length of the mate pair library is estimated incorrectly (Pop \textit{et al.}, 2004).

For the \textit{E.coli} data, SOPRA produces longest scaffolds that are also the most accurate as measured by genome coverage. The validity at 10 kb of these scaffolds is low because SOPRA’s estimate of the insert size is not accurate. For the \textit{C.elegans} data, MIP Scaffolder produces longest scaffolds that are almost as accurate as those produced by SOPRA. On this dataset, we note that scaffolds produced by SOPRA have a higher genome coverage than the contigs which is likely due to our method maximizing the scaffold coverage instead of genome coverage; contigs can be easier aligned to the same region in genome than scaffolds. For the \textit{P.syringae} data, MIP Scaffolder produces longest scaffolds that are not quite as accurate as those produced by the other methods. We also note that SSPACE performs better than SOPRA on this dataset. For the human scaffolds, the coverage figures are quite low for both MIP Scaffolder and SSPACE. This is partly due to the strict alignment...
criteria that was used to make running the validation feasible. Also the coverage of the contigs is low for this dataset indicating that a larger portion of the contigs are chimeric. A chimeric contig incorporated into a scaffold breaks the alignment of the scaffold against the reference into two parts of which only the longer one is considered in our validation method when computing genome and scaffold coverage. Structural variation between individuals may also cause lower coverage in the human data as the contigs are built on simulated reads from the reference genome, while the mate pairs are real data. Also validity at 50 kb is low indicating problems with scaffolding.

5 CONCLUSION

We have presented MIP Scaffolder which partitions the scaffolding problem into subproblems of restricted size and solves these subproblems exactly with mixed integer programming. We compared MIP Scaffolder to SOPRA and SSPACE on four datasets. On two of the datasets, MIP Scaffolder produced longer scaffolds that are not quite as accurate as those produced by the other two methods. For the human dataset, MIP Scaffolder produced slightly shorter but more accurate scaffolds than SSPACE and on the E.coli dataset SOPRA outperformed MIP Scaffolder. Our experiments also showed that our approach is fast allowing the user to try out different parameter combinations easily to optimize for long scaffolds. It is also possible to optimize for accurate scaffolds if, for example, EST data or the sequence of a close relative species can be used to measure the correctness of scaffolding.

ACKNOWLEDGEMENT

We wish to thank Rainer Lehtonen, Virpi Ahola, Ilkka Hanski, Panu Somervuo, Lars Paulin, Petri Auvinen, Lisa Holin, Patrik Koskinen and Pasi Rastas for insightful discussions about sequence assembly and scaffolding.

Funding: Academy of Finland [Grant numbers 118653 (ALGODAN) and 1140727]; Helsinki Graduate School in Computer Science and Engineering.

Conflict of Interest: none declared.

REFERENCES


