GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies

Alexey A. Gritsenko[1,2,3], Jurgen F. Nijkamp[1,3], Marcel J.T. Reinders[1,2,3] and Dick de Ridder[1,2,3]

[1]The Delft Bioinformatics Lab, Department of Mediamatics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
[2]Platform Green Synthetic Biology, P.O. Box 5057, 2600 GA Delft, The Netherlands
[3]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, The Netherlands

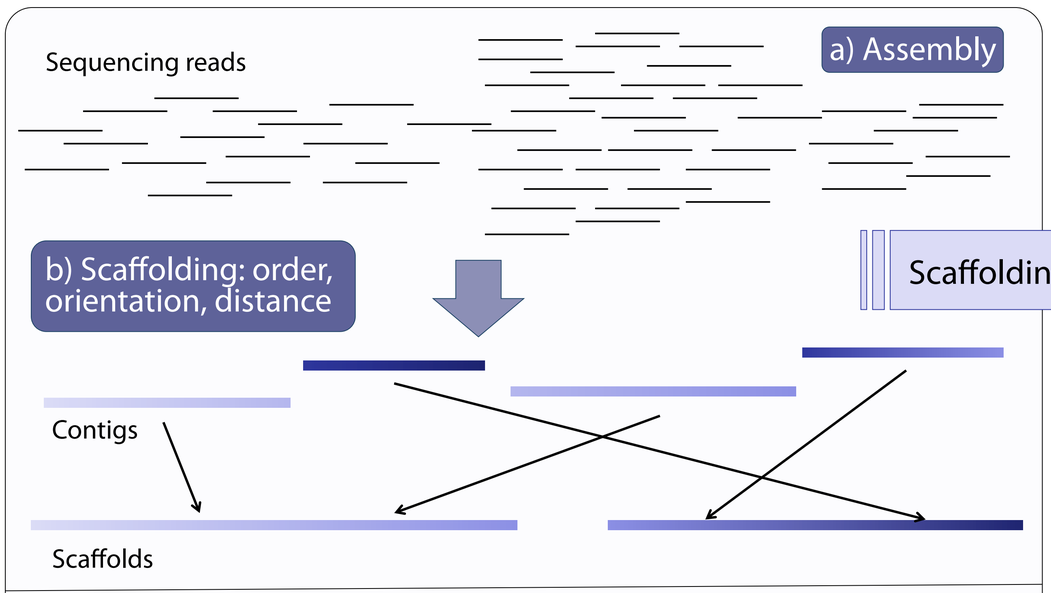

Fig. 1: HTS sequence assembly and scaffolding.

Introduction

High-throughput sequencing (HTS) technologies are increasingly used for *de novo* genome sequencing. The millions of short reads usually involved in HTS are first assembled into longer fragments called *contigs,* which are then scaffolded, i.e. ordered and oriented using additional information, to produce even longer sequences called *scaffolds* (Fig. 1). Most existing scaffolders can only use paired reads to perform scaffolding. We present GRASS (GeneRic ASsembly Scaffolder), which is an algorithm capable of using diverse *de novo* (i.e. paired reads) and comparative (i.e. related genomes) information sources for scaffolding.

Algorithm

The available scaffolding data is translated into weighted *contig links* l_j, as shown in step a of Fig. 3. The suggested relative order and orientation of contigs they connect, and approximate distance between them is deduced from the mapping (Fig. 2). Weights are chosen per data source.

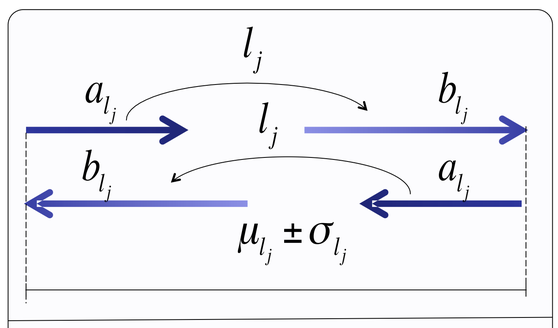

Fig. 2: Contig link constraints for equal orientation and forward order.

The links are used in an optimization approach (steps a-b), where we find an optimal position x_a and orientation t_a for each contig a, while trying to get as many of the links correct as possible. We can efficiently solve the resulting MIQP (*mixed-integer quadratic programming*) problem using an EM-like algorithm (step b). Finally, we can puzzle together the sequence by placing contigs according to the optimal x and t, and removing overlap when present (step c).

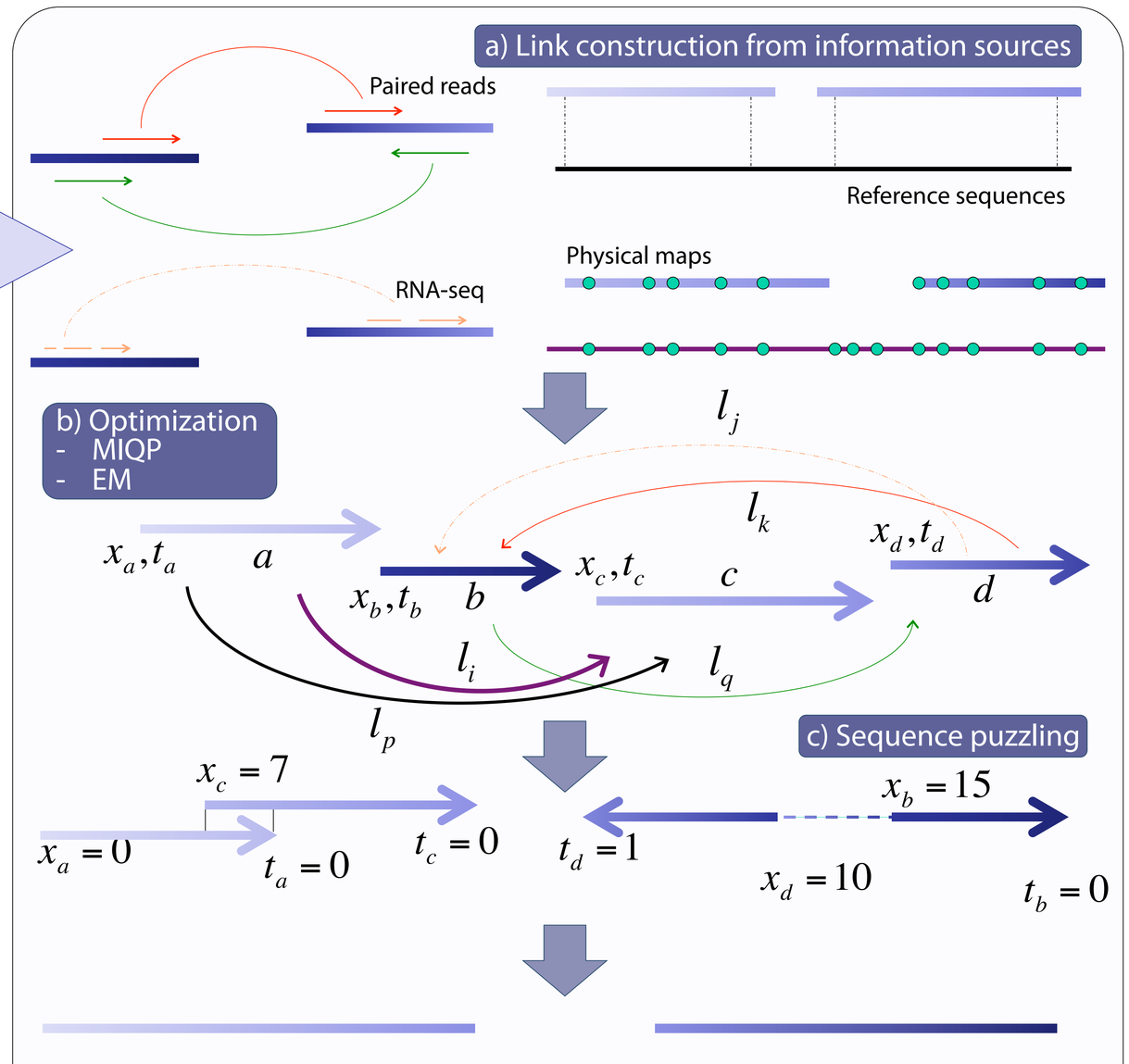

Fig. 3: GRASS algorithm overview.

Results

GRASS was compared to state-of-the-art scaffolders SSPACE, OPERA and the MIP Scaffolder on on *de novo* HTS assemblies (performed using Velvet) of three bacterial genomes: *Escherichia coli* K12, substr. MG1655; *Pseudoxanthomonas suwonensis* 11-1; and *Pseudomonas syringae* B728a. GRASS achieves a lower number of breakpoints while providing a competitive reduction in the number of contigs (Fig. 4). This result is further improved when genome sequences of the *E.coli* strains DH10B and BW2952 are used additionally to the reads to scaffold the MG1655 assembly (Fig. 4, left graph).

Conclusion

- We presented GRASS, a generic scaffolding algorithm suitable for combining multiple information sources. It achieves the best results when all available scaffolding information is used.
- GRASS constructed the most accurate scaffolds on the considered datasets.
- The accuracy/contiguity tradeoff displayed by GRASS puts it in a unique niche compared to existing scaffolders.

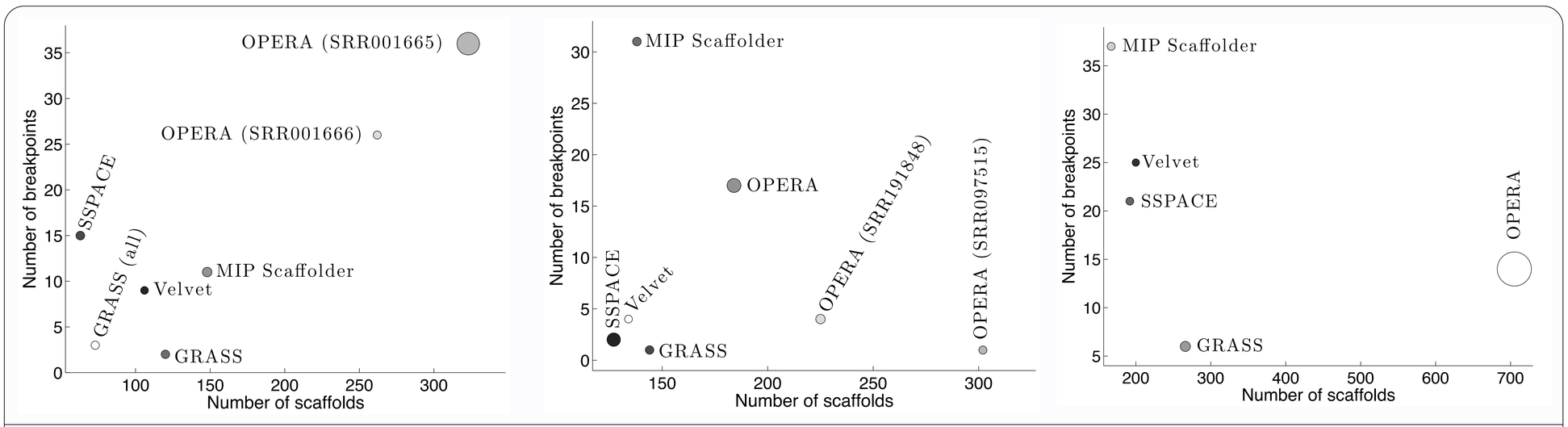

Fig. 4: Scaffolder comparison (from left to right) for *E.coli, P.suwonensis* and *P.syringae*. GRASS (all) additionally uses related genome data.

References

GRASS Gritsenko, A.A., Nijkamp, J.F., Reinders, M.J.T., de Ridder, D. GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies (*submitted*).

SSPACE Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W. (2011) Scaffolding pre-assembled contigs using SSPACE, *Bioinformatics,* 27, 578–579.

OPERA Gao, S., Sung, W.-K. and Nagarajan, N. (2011) Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences, *Journal of Computational Biology,* 18, 1681–1691.

MIP Scaffolder Salmela, L., Mäkinen, V., Välimäki, N., Ylinen, J. and Ukkonen, E. (2011) Fast scaffolding with small independent mixed integer programs, *Bioinformatics,* 27, 3259–3265.

Velvet Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Research,* 18, 821–829.

Kluyver CENTRE | Kluyver Centre for Genomics of Industrial Fermentation

PGSB Platform Green Synthetic Biology

The Delft Bioinformatics Lab

TU Delft Delft University of Technology