How to Reconstruct a Genome

Esko Ukkonen*

Department of Computer Science, University of Helsinki, Finland
Esko.Ukkonen@cs.helsinki.fi

Abstract. Since its early formulations (e.g., [1]), the genome assembly problem has attracted lots of interest from algorithm theoretic as well as from algorithm engineering point of view. In this problem, which is an inversion problem by nature, one is asked to reconstruct the entire DNA sequence from the short, randomly picked sequence fragments that a DNA sequencing instrument is able to read [2]. With the invent of current high-throughput sequencers producing such fragment reads in massive amounts, there is in molecular biology research a pronounced call for an accurate and fast reconstruction procedure.

It is customary to structure a reconstruction procedure into the following major steps: (1) Error correction of the fragments; (2) Finding pairwise overlaps between the fragments and representing the overlaps as a graph; (3) Constructing approximate superstrings, called *contigs*, for the fragments; (4) Constructing a linear order, called a *scaffold*, of the contigs. All steps are algorithmically challenging. Noisy data and intricate repetition structure of the target genome cause added difficulties.

The talk attempts to give an overall picture of the genome assembly process and its algorithmic aspects emphasizing some recent developments in error correction [3], contig assembly, and scaffolding [4]. We also try to convey experiences from a major undertaking of *de novo* sequencing of a higher organism, Glanville fritillary butterfly *Melitea cinxia*. (A collaboration with I. Hanski, www.helsinki.fi/science/metapop/index.htm).

References

1. Peltola, H., Söderlund, H., Tarhio, J., Ukkonen, E.: Algorithms for Some String Matching Problems Arising in Molecular Genetics. In: IFIP Congress, pp. 59–64 (1983)
2. Myers, G.: Whole-Genome DNA Sequencing. IEEE Computing in Science and Engineering 1, 33–43 (1999)
3. Salmela, L., Schröder, J.: Correcting errors in short reads by multiple alignments. Bioinformatics 27, 1455–1461 (2011)
4. Salmela, L., Mäkinen, V., Välimäki, N., Ylinen, J., Ukkonen, E.: Fast scaffolding with small independent mixed integer programs. Bioinformatics 27, 3259–3265 (2011)

* Supported by the Academy of Finland, grant 7523004 (Algorithmic Data Analysis).

B. Rovan, V. Sassone, and P. Widmayer (Eds.): MFCS 2012, LNCS 7464, p. 48, 2012.