DeBruijn graph assembly for optical maps

Steve Goldstein and David C. Schwartz

We report on a new *de novo* assembly algorithm for optical maps, Seed and Mature (SAM). The conceptual basis for SAM is an extension of the de Bruijn graph approach, the algorithm of choice for next-generation sequence assembly. Simply put, a whole genome optical map can be represented by the traversals of a certain graph and the assembly problem is to discover those traversals from the input data set, single-molecule restriction maps (*rMaps*). Specifically, we use geometric k-mer hashing [1] to identify nodes in the de Bruijn graph that are very likely error-free and then traverse the read paths implied by the rMaps containing instances of those nodes. This traversal allows us to localize the assembly; we use the optical map assembler Gentig[2] on the subsets of rMaps that are near each other on this graph. We call the resulting consensus maps *seed maps*. The seed maps typically cover most of the genome (*Medicago truncatula*: 85% coverage; N50 scaffold size 1.1 Mb) and they reliably approximate highly confident paths in the graph.

The seed maps are then extended and refined using an iterative scheme [3,4], producing another set of consensus maps (*Medicago truncatula*: 95% coverage; N50 scaffold size 22 Mb). The error rate for these consensus maps is sufficiently low so that we can construct the corresponding Euler path, assembling all but the most repetitive regions of the genome. We then attempt to fill gaps in the assembly by repeating the process, generating another set of seed maps and extending and refining them. For this set of seed maps, we use a lower stringency (smaller value of k) and use only those rMaps not already represented in the genome.

1. **Sequences, Maps, Genomes and Graphs: Graph Compression Algorithms for Efficiently Comparing Genomes.** S Goldstein, A Briska, S Zhou, and D Schwartz. UW Biostatistics and Medical Informatics Technical Report. (2004).

2. **Genomics via optical mapping III: Contiging genomic DNA and variations**. Anantharaman,T, Mishra,B, and Schwartz,DC. The Seventh International Conference on Intelligent Systems for Molecular Biology. 7:18-27.(1999).

3. **Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse.** Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, et al. PLoS Biol 7:5. (2009)

4. **High-resolution human genome structure by single molecule analysis.** B. Teague, M. Waterman,S. Goldstein,K. Potamousis,S. Zhou,S. Reslewic,D.Sarkar,A. Valouev,C. Churas,J. Kidd,S. Kohn,R. Runnheim,C. Lamers,D. Forrest,M. Newton,E. Eichler,M. Kent-First,U. Surti,M. Livny,D. Schwartz. PNAS. 107:24. (2010)

Author affiliation: Laboratory for Molecular and Computational Genomics, Genome Center, University of Wisconsin-Madison

Corresponding author: Steve Goldstein <sgoldstein@wisc.edu>