

ALLPATHS: A New Genome Assembly Algorithm that Preserves Intrinsic Ambiguity

Jonathan Butler,^{1,2} Michael Kleber,^{1,3} Pablo Alvarez,¹
 Will Brockman,¹ CheeWhye Chin,¹ Sante Gnerre,¹
 Manfred Grabherr,¹ Evan Mauceli,¹ Bruce Birren,¹ Eric S. Lander,¹
 David B. Jaffe¹

Keywords: whole genome assembly, polymorphism, k -mer, de Bruijn graph.

Introduction

Existing genome assembly algorithms like ARACHNE [1, 2] work well on prokaryotic genomes. For eukaryotic genomes, they correctly assemble well-behaved regions, but in general behave poorly on more complex regions embodying recent duplications or high polymorphism. To address these algorithmic defects, we are implementing a new general assembly algorithm, ALLPATHS, throughout which the inherent ambiguity of these regions is preserved.

Methodology

The algorithm works by first finding all high-quality sequence paths across each insert. The tractability of the search depends on replacing each sequence read with a path through the *de Bruijn graph on the set of k -mers* [3], in which each vertex represents a distinct k -mer in the genome, and a directed edge joining two vertices encodes a $(k + 1)$ -mer. The genome itself offers a natural compression scheme: we number each distinct k -mer appearing in the reads in such a way that consecutive kmers in the genome usually have consecutive numbers, i.e., k -mers i and $i + 1$ are usually connected by an edge. We can then replace sequence reads with the corresponding sequence of k -mer numbers and represent stretches of consecutively-numbered kmers as intervals, so an entire read may be represented internally as the sequence of intervals, e.g., [200, 300][290, 300][290, 450][700, 1000]. (Note that this read contains three copies of the eleven k -mer tandem repeat [290, 300].) We have found a k -mer numbering method that, while not optimal, sufficiently compresses the read paths to allow experimentation with full-coverage mammalian whole genome shotgun datasets, such as *Mus musculus* and *Canis familiaris*.

Since this k -mer numbering contains no information indicating how similar the sequence of one numbered k -mer may be to any other, we proceed with assembly by considering only perfect alignments. Any real data set contains errors, so we attempt to identify all low-quality k -mers in the reads and replace them with gaps of variable but bounded size. We have implemented a pairwise aligner for such gapped k -mer number strings (which we call k -mer paths). On top of this aligner, we have implemented a breadth-first search that attempts to walk from one end of an insert to the other using these alignments between k -mer paths.

Given sufficient high-quality coverage of an insert, the algorithm is mathematically guaranteed to find the true insert path, but may also find other (false) insert paths. ALLPATHS

¹Broad Institute at MIT and Harvard, 320 Charles Street, Cambridge, MA, 02141.

²E-mail: jbutler@broad.mit.edu

³E-mail: kleber@broad.mit.edu

will distinguish true from false insert paths by iteratively walking inserts (from small to large) and in so doing impose compatibility conditions between the paths for different inserts.

In areas of polymorphism, there will be multiple true paths, and, given a sufficiently recent and long duplication, false paths may remain even after the compatibility conditions are imposed. Therefore, at the end of the assembly process, where the large-scale genomic structure is obtained by piecing together insert paths, alternatives will be retained in those cases where they are intrinsic to the data.

Results

We implemented a prototype of the initial insert-walking algorithm and tested it on real data sets, for which high-quality coverage may be insufficient. For mouse, we found that if an insert path exists, then it agrees with finished sequence 95% of the time. In an initial analysis, we found that about half of the discrepant cases were attributable to finished sequence error and about half to algorithmic weakness. The inferred base error rate for paths themselves is about 1/80,000. At about the same rate, the paths corrected errors in finished sequence.

References

- [1] Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E. S. 2002. Arachne: A whole-genome shotgun assembler. *Genome Res.* **12**: 177–189.
- [2] Jaffe, D. B., Butler, J., Gnerre, S., Mauceli, E., Lindbad-Toh, K., Mesirov, J. P., Zody, M., and Lander, E. S. 2003. Whole-Genome Sequence Assembly for Mammalian Genomes: Arachne 2. *Genome Res.* **13**: 91–96.
- [3] Pevzner, P., Tang, H., and Waterman, M. 2001. An Eulerian path approach to DNA fragment assembly. *Proc. Nat. Acad. Sci.* **98**: 9748–9753