Poster: Scaffolding Draft Genomes using Paired Sequencing Data

J. Lindsay, J, Zhang, T. Farnham, Y. Wu, I. Mandoiu, R. O'Neill[1]
Computer Science & Engineering
[1]*Molecular and Cell Biology University of Connecticut*
Storrs, CT
james.lindsay@engr.uconn.edu

H. Salooti, E. Bullwinkel, A. Zelikovsky
Computer Science
Georgia State University
Atlanta, GA
alexz@cs.gsu.edu

The number of sequenced genomes is growing very quickly due to the low cost and availability of high throughput DNA sequencing platforms. However, most genome sequences are not complete, consisting of large numbers of contigs separated by gaps. The process of orienting and ordering these contigs, typically using pairs of reads with approximately known distance in the genome, is known as scaffolding. Scaffolding algorithms were first introduced along with the first genome assemblers. Much like the assemblers they were designed to work with pairs of relatively long Sanger reads (800-1000bp). The length of these reads ensures that the majority of them would map correctly onto contigs. Current sequencing platforms generate hundreds of millions of much shorter reads in each experiment. The shortness of the reads causes a large amount of non-unique and incorrect mapping. In this poster we present ongoing work on designing a scaffolding strategy appropriate for such type of data.

Our algorithm can scaffold contigs using paired-end or mate pair sequencing data from multiple platforms. The reads must first be mapped against the contigs, using any tool that reports multiple alignments for each read and can generate SAM output. Read pairs containing at least a read that is not uniquely mapped are removed from consideration. We also annotate contigs using RepeatMasker and RepeatModeler, and remove read pairs for which at least one read maps within an annotated repeat. Finally, read pairs consisting of reads that map in two different contigs are removed if the minimum insert size implied by the mapping is longer than the expected insert size by more than 3 standard deviations.

To orient the contigs, we first construct a graph $G = (V, E)$ with contigs as vertices and edges defined according to two redundancy parameters r and δ. To connect two contigs i and j by an edge we first require at least r read pairs with one read mapped onto contig i and the other mapped onto contig j. Read pairs mapped between i and j can either be consistent with current orientation of the two contigs, or consistent with switching the orientation of one of the contigs. The second constraint for adding an edge between u and v is that the ratio between the size of the larger set and that of the smaller one be larger than δ. For large values of δ (e.g., $\delta = 2$) we assume that the read pairs in the smaller set are incorrectly mapped and are thus implicitly discarded by using the above definition of edges. However, even with this definition, it may not be possible to find an orientation of the contigs that is consistent with all non-discarded read pairs. We use integer linear programming (ILP) to find the contig orientation that leaves the minimum number of inconsistent read pairs. The ILP uses 0/1 S_i to indicate the final orientation of each contig, with 1 indicating that the contig's orientation must be flipped. Additionally, we use 0/1 variables $S_{i,j}$ set to 0 if $S_i = S_j$ and to 1 otherwise. We denote by $h_{i,j}$ and $u_{i,j}$ the number of read pairs between contigs i and j that are consistent, respectively inconsistent with the initial contig orientations. This gives the following formulation:

$$\begin{array}{ll} \text{Min} & \sum_{i \in V} (h_{i,j} - u_{i,j}) S_{i,j} \\ \text{S.t.} & S_{i,j} \leq S_i + S_j,\ S_i + S_j + S_{i,j} \leq 2, \\ & S_{i,j} \geq S_j - S_i,\ S_{i,j} \geq S_i - S_j, \qquad \forall (i,j) \in E \end{array}$$

Experiments conducted using over 480 million 50bp SoliD mate pairs and contigs assembled from the published Sanger HuRef reads [1] show that our ILP produces more accurate contig orientations than the orientation tool included in the latest version of the BAMBUS scaffolding package [2].

Acknowledgments

This work has been supported in part by NSF awards IIS-0546457, IIS-0916401, IIS-0953563, and IIS-0916948.

References

[1] S. Levy *et al.*, "The diploid genome sequence of an individual human," *PLoS Biology*, vol. 5, no. 10, pp. e254+, 2007.

[2] M. Pop, D. S. Kosack, and S. L. Salzberg, "Hierarchical scaffolding with Bambus," *Genome Res.*, vol. 14, pp. 149–159, Jan 2004.

978-1-61284-852-5/11/$26.00 ©2011 IEEE