SILP3: Maximum Likelihood Approach to Scaffolding

Igor Mandric*, James Lindsay†, Ion Măndoiu ‡, and Alex Zelikovsky*
* Department of Computer Science, Georgia State University
Atlanta, GA 30302-3994,
email:{imandric, alexz}@cs.gsu.edu
† Computer Science and Engineering Department, University of Connecticut
Storrs, CT 06269
email:{james.lindsay, ion}@engr.uconn.edu

***Abstract*—Scaffolding is the important stage of genome assembly consisting of orienting and ordering contigs based on read pairs. We present a scalable scaffolding algorithm that finds most likely contig orientation using integer linear program solved by a non-serial dynamic programming approach. We then formulate the problem of finding most likely contig ordering as an optimization problem and propose a novel ordering algorithm improving over our previous scaffolding tool SILP2.**

I. Introduction

Next generation sequencing has become a powerful technology broadly used in genome assembly. The input for the scaffolding problem represents the set of contigs that were obtained from the previous stages in the genome assembly pipeline together with the read pairs that provide linkage information between the contigs. Scaffolding provides an orientation of each contig, a linear order and estimated gaps between adjacent contigs in each scaffold (each scaffold is a chain of contigs). Our approach is to look for a scaffolding of maximum likelihood.

The flow of SILP2 [3] consists of the following steps:

1) mapping reads onto contigs
2) scaffolding graph construction
3) maximum likelihood contig orientation via ILP
4) decomposition into paths of orientation compatible edges via bipartite matching
5) maximum likelihood gap estimation

II. Maximum Likelihood Orientation

Let $G = (V, E)$ be the scaffolding graph connecting vertices-contigs with edges-read pairs. Maximum likelihood contig orientation can be formulated as the following ILP [3]. Let S_i be a boolean variable with value being set to 0 if the orientation of contig i remains unchanged. We identify 4 states A, B, C, D in which a pair (i,j) of contigs can be based on their orientation and relative ordering. We introduce 4 boolean variables $A_{ij}, B_{ij}, C_{ij}, D_{ij} = \{0,1\}\ \forall(i,j)$. For each state a weight A^w_{ij}, B^w_{ij}, C^w_{ij}, D^w_{ij} using a maximum likelihood approach is calculated. The number of concordant contig pairs is then maximized:

$$Max \sum_{(i,j)\in E} A^w_{ij} \cdot A_{ij} + B^w_{ij} \cdot B_{ij} + C^w_{ij} \cdot C_{ij} + D^w_{ij} \cdot D_{ij}$$

subject to constraints connecting A_{ij}, B_{ij}, C_{ij}, D_{ij} with S_i. We apply the technique of non-serial dynamic programming in order to solve the optimization problem. The output of this step is a directed graph $G' = (V, E', w, g)$ in which each contig and each edge has the most likely orientation, $w : E' \to R^+$ is the weight of an edge contributing to the ILP objective, and $g : E' \to R^+$ is the estimated gap between adjacent contigs.

III. Maximum Likelihood Ordering

An *ordering* O is a graph consisting of a set of disjoint directed chains of contigs together with estimations of gaps between all pairs of adjacent contigs. Note that in the ordering O the estimation of gaps between adjacent contigs uniquely defines the gap $g_O(i,j)$ between any pair of connected contigs i and j. SILP2 [3] extracts the maximum subgraph-ordering out of G'. Instead, SILP3 is aimed to find an ordering with the maximum support of G'-edges. A directed edge $e = (i,j) \in E'$ is *concordant* with an ordering O if i precedes j in O and the gap $|g(e) - g_O(i)| \leq 5\sigma$, where σ is the standard deviation of read pair fragment length. The maximum likelihood ordering is approximated with the ordering concordant with the edges of G' of the maximum total w-weight.

We first run DFS on G' recursively deleting least-weight edge resulting in G' being a DAG. Any topological order of G' is order-concordant with all edges remaining in G'. We are constructing a topological order in the following way. Let C be the set of contigs between i and j in G'. We sort C based on their distance estimations to i and j, $C_{sorted} = \{i_1, i_2, ..., i_k\}$. All edges in the graph are replaced with the directed path $P_0 = \{i, i_1, i_2, ..., i_k, j\}$. Then we determine the maximum weight bipartite matching similar to SILP2.

IV. Results

Scaffolding is a challenging and a very important part of the de novo assembly pipeline. We assess the quality of SILP3 based on the metrics presented in [1]. SILP3 outperforms SILP2 on accuracy and it is competitive to ther scaffolders.

References

[1] M. Hunt et al. A comprehensive evaluation of assembly scaffolding tools. Genome Biology, 15(3), 2014.

[2] O. Shcherbina. Nonserial dynamic programming and tree decomposition in discrete optimization. In OR, pages 155-160, 2006.

[3] J. Lindsay, H. Salooti, A. Zelikovsky, and I. Măndoiu. 2012. Scalable genome scaffolding using integer linear programming. *In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB '12).* ACM, New York, NY, USA, 377-383.

978-1-4673-1321-6/12/$31.00 2014 IEEE