Computational methods for Multiple Genome Alignment and Synteny detection

Jagan Jayaraj

Abstract:

Multiple genome alignments are useful to detect synteny, gene order, and large-scale genomic re-arrangements which help to understand genome evolution, divergence and the development of protein functions. However, aligning multiple whole genomes is very computationally intensive [3] and many genomes are only partially complete. Fast approximation algorithms have been developed to handle both the issues. Also, some visualization frameworks have been developed to interpret the results. This report is a survey of algorithms and tools for multiple genome alignment.

Introduction:

Whole genome comparisons are considered as solution to identify coding regions, regulatory signals and deducing the mechanisms and history of genome evolution. Alignment is a part of comparative genomics based on collinear arrangement of sequence similarities [2]. However, some consider even sequence re-arrangements as part of the alignment problem [4]. Multiple genome alignment is multiple sequence alignment performed over the entire genomes. The deluge of sequence data in the last few years has resulted in the creation of multiple genome alignment algorithms. In-fact, the first true multiple genome alignment tool, MGA, was developed in 2002 [3]. Since then, many algorithms have been developed for closely-related species but some algorithms do exist for distantly-related organisms [4].

Large sequence alignments can be approached in two ways: (1) iterative pair-wise alignment (2) anchor-based multiple alignment. Approach one is an ideal extension of standard dynamic programming approach but because of time complexity it is impractical. Anchor-based approaches identify substrings that are very likely part of global multiple alignment, align them and attempt to fill the gaps [3]. Methods vary in the way they compute anchors and how they align them. Except for a few exceptions, all the methods presented here are anchor-based.

This report discusses the evolution of multiple genome alignment algorithms from pair-wise alignment algorithms to more sophisticated alignment algorithms along with the visualization tools. The first section talks about first-generation pair-wise alignment algorithms, the second on true multiple genome aligners, and the third on the visualization tools. A good survey on multiple genome aligners prior to 2004 is available at [2]. Survey on comparative genomic tools for prokaryotes can be found in [1]

First-generation pair-wise alignment algorithms:

Many methods used a generate-and-test approach: k-mers/'seed's are generated based on say hashing and tested to see if they could be extended to a larger contiguous match called anchor. Extension of the k-mers is done by pair-wise comparison and the method is plagued by long runtimes. DIALIGN constructs 'gap-free' fragments and aligns them using dynamic programming. Based on hashing, ASSIRC comes with k-mers and applies graph-theoretic algorithms to extend them. PipMaker is based on a high-performance

local alignment algorithm called BlastZ, variant of gapped BLAST, and masks repetitive regions to speed computation [2]. GLASS is designed for eukaryotic organisms with strongly conserved exons but suffers from large space requirements [3] [2]. WABA treats the *wobble base,* the third base in a codon, differently and used HMM to align homologous regions. LSH-ALL-PAIRS uses a randomized search technique called locality-sensitive hashing to improve the speed of seed matching. It beats the other methods like GLASS, ASSIRC and PipMaker in its genre [2]. MUMmer is the first algorithm to use suffix-trees, a data-structure, to speed up anchor detection. MUMmer 2.0 improved the computation time by not constructing suffix-tree for one of the sequences. MUMmer suffers from the bigger size of the suffix tree and its inability to accommodate Multiple Exact Matches in a well-defined manner [3]. NUCmer and PROmer alignment algorithms were developed on top of MUMmer to support inexact matches. Vmatch uses enhanced suffix arrays to save space and improve time. It doesn't extend k-mers but instead computes them directly. Vmatch is better computationally than all the methods discussed above. LAGAN computes anchors and performs Needleman-Wunsch on a pruned search space [5].

True multiple genome alignment algorithms:

Multiple Genome Aligner (MGA) was the first tool to align more than two whole genome sequences. It employs enhanced suffix arrays to find anchors, maximal multiple exact matches (multiMEMs), in linear time and space. Dynamic programming is used to find the longest non-overlapping multiMEMs and a greedy approach to fill the gaps [3].

MUMmer 3.0 is open-source, has a very efficient implementation of suffix trees and supports MEMs [13]. Multi-LAGAN is a progressive aligner built on top of LAGAN [5]. MAVID is a progressive alignment approach based on generating maximum likelihood ancestor sequence given any two sequences. If no phylogenetic tree is given, it starts with an initial random tree and iteratively refines the tree. Gene predictors are used to generate anchors for pair-wise alignment [6]. DIALIGN P is a parallel multiple genome aligner [12].

Mauve handles large-scale evolutionary processes by supporting re-arrangement of locally collinear blocks (LCB – a bunch of anchors). It does recursive anchoring and gapped alignments to fill the gaps [4]. EMAGEN uses enhanced suffix arrays, graph theory concepts, and CLUSTAL W for gaps [7]. GAME is a generate-and-test algorithm for MEM anchors but using effective MEM filtering techniques to speed up the computation [8]. [10] uses novel chaining strategy based on dynamic programming and uses progressive multiple alignment to close the gap. MULAN handles both draft and high-quality sequences. It is based on TBA aligner and outputs local multiple alignments [11]. SMGA is the latest and the authors claim is the fastest [9].

Visualization tools:

Visualization tools are equally important as the algorithms to do multiple genome alignments. Many visualization tools have been developed concomitant with the alignment tools and lack versatility [2].

PIP displays PipMaker's alignments and is tuned to highlight synteny. Alfresco is an interactive front-end to display alignments from multiple sources. Intronerator is developed for WABA alignments and is specific to nematode community. VISTA works with both GLASS and MUMmer. SynPlot is developed for DIALIGN and like the tools mentioned above displays only collinear segments, and doesn't display draft sequences. DisplayMUMs was designed primarily for examining sequence assembly and visualizing polymorphisms [2].

MGA has an option to write out XML files [3]. MapView was created for MUMmer 3.0. Latest tools like MAUVE, Multi-LAGAN, MULAN, MAVID and GAME come with in-built visualizers.

Conclusion:

Necessity for multiple genome alignments tools arose very recently but the tools have matured a lot over few years. From doing just pair-wise alignments, the tools have progressed to align hundreds of genomes very efficiently. This report covered some of the latest multiple genome alignment and visualization tools. As alignment tools arose in response to a need for automation, these tools could be incorporated into bigger sequence and automation frameworks.

APPENDIX: AVAILABILITY OF ALGORITHMS AND VIEWERS

Alignment algorithms	
ASSIRC	Program only: ftp://ftp.biologie.ens.fr/pub/molbio/
DIALIGN	Program and server: http://bibiserv.TechFak.Uni-Bielefeld.DE/dialign/
MUMmer	Program only: http://www.tigr.org/software/mummer/
PipMaker/BlastZ	Server only: http://bio.cse.psu.edu/pipmaker/
GLASS	Program and server: http://crossspecies.lcs.mit.edu/
WABA	Program and server: http://www.soe.ucsc.edu/~kent/xenoAli/or http://www.cse.ucsc.edu/~kent/xenoAli/
LSH-ALL-PAIRS	Not available on the Internet, must contact the author at: jbuhler@cs.washington.edu
Vmatch	http://www.vmatch.de
MGA	http://bibiserv.techfak.uni-bielefeld.de/mga/
Comparative alignment viewers	
PipMaker/BlastZ	Server only: http://bio.cse.psu.edu/pipmaker/
Alfresco	Program and server: http://www.sanger.ac.uk/Software/Alfresco/
Intronerator	Server only: http://www.cse.ucsc.edu/~kent/intronerator/
VISTA	Program and server: http://www-gsd.lbl.gov/vista/
SynPlot	Program only: http://www.sanger.ac.uk/Users/jgrg/SynPlot/
ACT	Program only: http://www.sanger.ac.uk/Software/ACT/
DisplayMUMS	Program only: http://www.tigr.org/software/displaymums/

Figure 1: Taken from [2].

References:

1) Dawn Field, Edward J. Feil and Gareth A. Wilson; Databases and software for the comparison of prokaryotic genomes.

2) Chain P, Kurtz S, Ohlebusch E, Slezak T: An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. Brief Bioinform 2003, **4:**105-123.

3) Hohl M, Kurtz S, Ohlebusch E.; Efficient multiple genome alignment; Bioinformatics 2002

4) Darling et al.; Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements; Genome Research 2004

5) Brudno M, Do CB, Cooper GM, Kim MF, Davydov E; NISC Comparative Sequencing Program; Green ED, Sidow A, Batzoglou S; LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA; Genome Res. 2003 Apr;13(4):721-31.

6) Bray et al; MAVID: Constrained Ancestral Alignment of Multiple Sequences; Genome research 2004

7) Deogun et al.; EMAGEN: An Efficient Approach to Multiple Whole Genome Alignment;

8) Choi et al; GAME: A simple and efficient whole genome alignment method using maximal exact match filtering

9) Hu et al; A Fast Algorithm Aligning Multiple Microbial Genomic Sequences;Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference

10) Ma et al, New Model for Global Multiple Alignment of Whole Genome Sequences; International Journal of Information Technology Vol. 11, No. 8, 2005

11) Ovcharenko I et al.; Mulan: multiple-sequence local alignment and visualization for studying function and evolution; Genome Res. 2005

12) Martin Schmollinger, Kay Nieselt, Michael Kaufmann and Burkhard Morgenstern; DIALIGN P: Fast pair-wise and multiple sequence alignment using parallel processors; *BMC Bioinformatics* 2004, 5:128

13) Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg ;Versatile and open software for comparing large genomes; Genome Biology 2004.

14) Jeong-Hyeon Choi, Kwangmin Choi, Hwan-Gue Cho and Sun Kim; Multiple Genome Alignment by Clustering Pairwise Matches