CIRCALIGN: A new algorithm to align multiple mitochondrial genomes

Francisco Fernandes,[1] Luísa Pereira,[2] and Ana T. Freitas[1]

1 INESC-ID/IST, Lisboa
2 IPATIMUP, Porto

The comparison of homologous sequences from different species is an essential approach to reconstruct the evolutionary history of species and of the genes they harbour in their genomes. Several complete mitochondrial and nuclear genomes are now available, increasing the importance of using multiple sequence alignment algorithms in comparative genomics. In this paper we propose a new and efficient anchor-based multiple sequence alignment algorithm that can deal with circular genomes. The algorithm includes de CSA algorithm, recently published, to identify the best rotation for a set of circular genome sequences that are to be aligned. For the multiple sequence alignment a set of anchors is found by recursively calculating the largest chain of common blocks from the current region in all the sequences, and then the gaps between the fixed blocks are aligned using a variation of the Needleman-Wunsch global alignment algorithm using a dynamic score function.

I. MOTIVATION:

Genomic multiple sequence alignment tools have been playing an important role in comparative genomics and phylogenetic reconstruction. Despite the fact that these tools are heuristic based and sometimes lead to poor biologically plausible alignments, they were also developed to deal only with linear genomic sequences. When applied to circular genomes, the results become extremely sensitive to the exact place where the genomic sequence begins. This limitation is very important since circular DNA sequence alignments are central to a number of biological problems.

Mitochondrial DNA (mtDNA) has long been used for phylogenetic analyses. In fact, the absence of recombination in this genome enables an easy and direct inference of the phylogenetic evolution and its fast mutation rate leads to a high discriminative power. Until recently, phylogenetic reconstructions were based on certain regions of the mtDNA molecule, mainly the protein-coding gene cytochrome b when comparing different species [1] and hypervariable regions on D-loop when comparing human populations (e.g. [2]). But the high recent throughput of automatic sequencing techniques is offering the possibility to study complete mtDNA genomes in humans (see revision in [3]) and in other species (ex: Mus musculus, [4]). By the end of April 2009 there were around 5,650 human mtDNA complete genomes in GenBank [5] and 1,750 complete mtDNA which should be used as reference sequence for the diverse species in RefSeq [6]. The blind application of standard phylogenetic analyses in these massive datasets without concern to the circularity of these molecules will lead to the overestimation of genetic distances between species.

II. BACKGROUND:

Traditional genome sequence alignment algorithms based on dynamic programming are very inefficient when multiple long genome sequences needed to be aligned. To tackle this problem several heuristic based methods have been proposed. The most popular progressive multiple sequence alignment (MSA) method is ClustalW [7], to which access is provided by a number of web portals. Other methods like T-COFFEE [8], DIALIGN [9], MUSCLE [10], MLAGAN [11], MAVID [12], and MAUVE [13] are also widely used. However since all these tools are heuristic based they are not always successful in constructing biologically plausible alignments. Despite this fact, what really matters is that they are commonly used to align all type of genomic sequences, including sequences that correspond to circular genomes. Since these tools were

developed to deal with linear genomic sequences, when applied to circular genomes the results become very sensitive to the exact place where the genomic sequence begins. This is an important limitation, because circular genome alignment is central to a number of biological problems.

Two software tools have been recently proposed to align circular DNA genome sequences: Circal [14] and Cyclope [15]. The algorithm implemented in the Circal package uses a complex gap cost function and can only deal with short sequences, less than a thousand characters, due to its time complexity. The Cyclope package includes an implementation of an exact and a heuristic method with time complexities that are prohibitive if it is to be used to align several sequences with several thousands of base pairs.

In this paper, we present a new MSA algorithm as an extension to CSA [16], a very efficient algorithm that finds the correct rotation for a set of circular genome sequences that are to be aligned. Firstly, the genomic sequences are circularized, and then the best rotation is calculated based on the largest chain of non-repeated blocks that belongs to all the sequences. This chain is obtained with the help of a generalized cyclic suffix tree, which was a new concept introduced in that work [16]. At the end of the process, the users can visualize all the identified common blocks, obtaining a precise idea on how these regions are conserved along the genomic sequences.

III. METHODS:

After the rotation phase using the CSA algorithm, the CIRCALIGN tool aligns the optimally rotated sequences using an anchor-based multiple sequence alignment algorithm, similar to MAVID [12] or MAUVE [13]. For alignment purposes, CIRCALIGN uses Multiple Maximal Exact Matches (Multi-MEMs) as anchors, which may occur more than once on each sequence. A Multi-MEM instance is defined by its starting positions on each sequence and by its length, or “weight”. Two Multi-MEMs are said to be collinear if the ending positions of the first Multi-MEM occur before the starting positions of the second Multi-MEM on every sequence. CIRCALIGN considers only Multi-MEMs occurring an equal number of times in the current region of each sequence. It efficiently retrieves all the Multi-MEMs from the single already built “global” suffix tree. This allows to scan the suffix tree only once, and dynamically sort and effectively reuse prefixes and suffixes of Multi-MEMs that are relevant in one region but not in another.

Firstly CIRCALIGN “cleans up” the cyclic suffix tree built by the CSA algorithm in order to convert it into a regular “linear” suffix tree, then the Multi-MEMs and their corresponding positions on each sequence are extracted. The optimal chain of these Multi-MEMs is then calculated using the Heaviest Increasing Subsequence algorithm [17], implemented over a priority queue. This chain defines the set of anchors for the current region. The same procedure is recursively applied to the gaps between the previous fixed anchors until there are no more Multi-MEMs for that region or when the region has zero length on at least one of the sequences. Anchors that unluckily matched on obvious wrong location and could lead to bad alignments are also automatically discarded by detecting discrepancies in the consistency of the size of the gaps around that anchor on each sequence. The unclosed remaining gaps are aligned using a progressive alignment approach based on a dynamic programming algorithm similar to Needleman-Wunsch [18]. The subsequences inside the gaps are aligned in pairs following the increasing order of their lengths.

The dynamic programming stage does not take into account the current consensus sequence, but it uses a new approach by keeping an expanding array of the size of the current alignment consensus with the information about the frequency of each alphabet symbol on each consensus position. So, in fact, each new sequence is not aligned against a consensus sequence, but is, in some sense, aligned to every other previously aligned sequence. For this alignment a dynamic score is computed by looking at the symbol in the current sequence and the number of already aligned equal symbols in that consensus position of all the previous aligned sequences. When aligning the K-th sequence of a set of N sequences, the dynamic programming recurrence is given by the following formula:

$$M[i][j] = \max\begin{cases} M[i-1][j-1] + \sum_{l=1}^{|\Sigma|}\left(C(\alpha_l, j) * S(s_i^k, \alpha_l)\right) & , diagonal \\ M[i-1][j] + (k-1) * S(s_i^k, '-') & , up \\ M[i][j-1] + \sum_{l=1}^{|\Sigma|}\left(C(\alpha_l, j) * S('-', \alpha_l)\right) & , left \end{cases}$$

where $\sum$={A,C,G,T,-} is the alphabet, α_l is the l-th symbol of the alphabet, s^k_i is the i-th symbol of the k-th sequence, C(α,j) is the number of occurrences of symbol α in the j-th column of the consensus, and S(α,β) is the substitution/match score when aligning symbol α to symbol β. With N=2, when aligning only two sequences, the formula trivially reduces itself to the regular Smith-Waterman recurrence.

IV. RESULTS:

In order to evaluate the efficiency of the proposed tool when aligning multiple circular genomes and the biological relevance of the obtained alignments, we conducted tests using three sets of mtDNA sequences. The first set includes sequences of 16 Primates, the second set includes sequences of 12 Mammals and the last set is a set of distantly related sequences including 19 mtDNA sequences (the 16 Primates, the *Drosophila melanogaster*, the *Gallus gallus* and the *Crocodylus niloticus*). Due to space limitations, datasets details were included as supplementary material at http://kdbio.inesc-id.pt/~fjdf/circalign/SequenceNamesTable.doc

Multiple sequence alignments were performed using the ClustalW [7] tool and the new CIRCALIGN algorithm. Alignment quality was compared in two ways: (1) by evaluating the number of conserved columns and the correspondent sum-of-pairs score (see Table 1); (2) by evaluating genetic standard measures in the software Arlequin [19] (see Table 2).

Table 1 – Comparison between ClustalW and CIRCALIGN.

	Mammals (12 sequences)		Primates (16 sequences)		Set3 (19 sequences)	
	ClustalW	CIRCALIGN	ClustalW	CIRCALIGN	ClustalW	CIRCALIGN
Consensus size	18612	20612	17447	19271	20180	22531
Average gaps per sequence	1826	3826	865	2689	3421	5772
Number of conserved columns	6617	7401	7152	7744	3119	4627
Sum-of-Pairs score	425123	453932	1034332	1056411	1035639	1169318
Running Time	53m56s	0m25s	1h27m0s	0m29s	1h53m45s	5m35s

Regarding the results presented in Table 1, it is possible to see that the CIRCALIGN alignment algorithm outperforms the reference tool ClustalW in both speed and accuracy in terms of the number of totally conserved columns and the sum-of-pairs score in all the performed tests. These encouraging results in the scoring measures were achieved with the new dynamic score that instead of considering a consensus sequence, maintains for each position the count of symbols of all the previously aligned sequences.

From a biological point of view (see results in Table 2), CIRCALIGN reduces the number of substitutions, both transitions and transversions, leading to lower nucleotide diversities and mean number of pairwise differences between sequences. Those lower diversity measures result in an increase in the number of gaps, in the regions between the aligned blocks, conducting also to an increase in the consensus size. These gaps are, most probably, regions of substitutions, being the blocks usually separated by one, two or three gaps in the coding region, reflecting the trinucleotide codon structure, and by more gaps in the non-coding region, where the functional

constraints are weaker. In accordance with this supposition, the ratio for gaps between ClustalW and CIRCALIGN is about one half, justified by ClustalW having high penalties for gaps. Curiously, that ratio is maintained along the three sets of sequences (0.44 for Primates, 0.48 for Mammals and 0.47 for the Primates with three distant species).

Table 2 – Genetic diversity standard measures for Primates, Mammals and Primates with more distantly related sequences, aligned with both alignment tools.

		ClustalW	CIRCALIGN
First set (Primates)	size (bp)	18033	19271
	polymorphic sites	10910	11527
	transitions	9429	7324
	transversions	5226	4040
	substitutions	14655	11364
	indels	2544	5780
	Mean no. of pairwise differences	4303 +/- 1939	4263 +/- 1921
	Nucleotide diversity	0.239 +/- 0.120	0.221 +/- 0.112
Second set (Mammals)	size (bp)	19220	20612
	polymorphic sites	12591	13211
	transitions	9987	6805
	transversions	6625	4988
	substitutions	16612	11793
	indels	3916	8122
	Mean no. of pairwise differences	5640 +/- 2592	5591 +/- 2570
	Nucleotide diversity	0.293 +/- 0.152	0.271 +/- 0.140
Third set (Primates + Drosophila melanogaster + Gallus gallus + Crocodylus niloticus)	size (bp)	19964	22531
	polymorphic sites	17514	17904
	transitions	13594	9173
	transversions	9777	6487
	substitutions	23371	15660
	indels	5473	11641
	Mean no. of pairwise differences	5892 +/- 2631	5557 +/- 2481
	Nucleotide diversity	0.295 +/- 0.147	0.247 +/- 0.123

V. CONCLUSIONS:

As stated before the new proposed algorithm to align multiple circular genomes outperforms the reference tool ClustalW in a set of important aspects. Despite the fact that the results obtained are encouraging, the inclusion of a large number of gaps in the final alignment can raise important biologically relevant issues. Currently, a set of new approaches are under consideration in order to overcome this problem and obtain an optimal balance between the number of matching symbols and the number of included gaps.

It is important to notice that with the CIRCALIGN tool the maximum speedup improvement achieved was of 180x times faster when compared with ClustalW. This performance improvement is mainly due to the efficient use of anchors that allows to process only the gaps between those anchors, which greatly reduces the computation time of the dynamic programming part.

This new multiple alignment algorithm is planned to be released as a standalone general purpose multiple sequence alignment tool in a near future.

VI. ACKNOWLEDGEMENTS:

This project was supported by the ARN project (PTDC/EIA/67722/2006) from FCT. FCT partially supports IPATIMUP through POCTI, Quadro Comunitário de Apoio III.

VII. REFERENCES:

1. Castresana J: Cytochrome b phylogeny and the taxonomy of great apes and mammals. Mol Biol Evol. 2001, 18:465-471.
2. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Gölge M, Dimitrov D, Hill E, Bradley D, Romano V, Calì F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Nørby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ: Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet. 2000, 67:1251-1276.
3. Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, Costa S, Máximo V, Macaulay V, Rocha R, Samuels DC: The diversity present in 5,140 human mitochondrial genomes. Am J Hum Genet. 2009, 84:628-640.
4. Goios A, Pereira L, Bogue M, Macaulay V, Amorim A: mtDNA phylogeny and evolution of laboratory mouse strains. Genome Res. 2007, 17:293-298
5. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: GenBank. Nucleic Acids Res. 2008, 36(Database issue):D25-30.
6. Pruitt KD, Tatusova T, Maglott DR: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007, 35(Database issue):D61-65.
7. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: Clustal W and Clustal X version 2.0. Bioinformatics, 2007, 23:2947-2948.
8. Notredame C, Higgins DG, Heringa J: T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000, 302(1):205-17.
9. Brudno M, Chapman M, Göttgens B, Batzoglou S, Morgenstern B: Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinformatics 2003, 4:66.
10. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004, 32(5):1792-1797.
11. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 2003, 13(4):721-731.
12. Bray N, Pachter L: MAVID: constrained ancestral alignment of multiple sequences. Genome Res. 2004, 14:693-699.
13. Darling AC, Mau B, Blattner FR, Perna NT: Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004, 14:1394-1403.
14. Fritzsch G, Schlegel M, Stadler PF: Alignments of Mitochondrial Genome Arrangements: Applications to Metazoan Phylogeny. J Theor Biol 2006, 240:511-520.
15. Mosig A, Hofacker IL, Stadler PF: Comparative Analysis of Cyclic Sequences: Viroids and other Small Circular RNAs. In Proceedings of the German Conference on Bioinformatics: 20-22 September 2006; Tübingen. P-83:93-102.
16. Fernandes F, Pereira L, Freitas AT: CSA - An efficient algorithm to improve circular DNA multiple alignment. BMC Bioinformatics 2009, 10:230.
17. Jacobson G, Vo K-P: Heaviest increasing/common subsequence problems. Proc. 3rd Annual Symposium on Combinatorial Pattern Matching, volume 644 of Lecture Notes in Computer Science, pages 52-66, Berlin. 1992. Springer-Verlag.
18. Needleman SB, Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. 1970, J Mol Biol 48 (3): 443–53.
19. Excoffier L, Laval G, Schneider S: Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 2005, 1:47-50.