**MASC:**

**A Linear Method for Multiple Nucleotide Sequence Alignment on Spark Parallel Framework**

Wenhe SU, School of Computer Science and Technology, National University of Defense Technology, Changsha, China

Quan ZOU, School of Computer Science and Technology, Tianjin University, Tianjin, China

Xiangke LIAO, School of Computer Science and Technology, National University of Defense Technology, Changsha, China

Yutong LU, School of Computer Science and Technology, National University of Defense Technology, Changsha, China

Shaoliang PENG\*, School of Computer Science and Technology, National University of Defense Technology, Changsha, China

\*Corresponding author: E-mail: pengshaoliang@nudt.edu.cn

Connect Information:

|  |  |  |
| --- | --- | --- |
| Name | Telephone | E-mail |
| Wenhe SU | +86 13272094630 | suwenhecn@gmail.com |
| Quan ZOU | +86 17092261008 | zouquan@nclab.net |
| Xiangke LIAO | +0731 84576322 | xkliao@nudt.edu.cn |
| Yutong LU | +0731 84576322 | ytlu@nudt.edu.cn |
| Shaoliang PENG | +86 13574817196 | pengshaoliang@nudt.edu.cn |

**Abstract**

**Multiple sequence alignment (MSA) is an essential prerequisite and dominant method to deduce the biological facts from a set of molecular biological sequences, including homology modeling, secondary structure prediction, phylogenetic reconstruction, and protein structure and function prediction. In this work, we take advantage of a center-star strategy to reduce the MSA problem to pairwise alignments, and we use a suffix tree to match identical substrings between the two pairwise sequences. We can accomplish MSA in O(*mn*) time in this manner, where *m* is the number of sequences, and *n* is the average length of the sequences. Furthermore, we execute our method on the Spark distributed parallel framework to deal with ever-increasing massive datasets. Our method is significantly faster than previous techniques, with no loss in accuracy for highly similar nucleotide sequences like homologous sequences, which we experimentally demonstrate. This work will facilitate 16S rRNA metagenomics analyses and other research.这句话是摘要的conclusion 审稿人认为太weak了。**

**Key words:** Multiple Sequence Alignment, Suffix Tree, Center-Star Strategy, Spark

1. **Introduction**

MSA is the alignment of more than two molecular biological sequences, aiming to discover maximally similar (or identical) amino acid or identical nitrogenous base positions across the aligned query set of sequences. Multiple sequence alignment (MSA) is an essential preprocess to many bioinformatics analyses, including homology modeling, secondary structure prediction, phylogenetic reconstruction, and protein structure and function prediction. An MSA is visualized as a two-dimensional matrix in which the rows are the individual sequences and the columns are maximally similar or identical amino acid or nitrogenous base positions arranged to correspond by inserting gap characters in appropriate positions (known as indels). An MSA can provide a wealth of information about the structure/function relationships within a set of sequences, such as the evolutionary conservation of functionally or structurally important sites, and conserved hydrophobicity patterns in precise regions.

Although dynamic programming using the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970) can be generalized in theory to produce an alignment for any number of sequences, unfortunately, this generalization leads to an explosive increase in computer time and memory requirements as the number of sequences increases (Taylor, 1990). MSA remains under continuous development, and is regarded as one of the most challenging problems in the field of bioinformatics and computational biology. (Chatzou *et al.*, 2016) Furthermore, the computation of an accurate MSA has long been known to be an NP-complete problem (Wang and Jiang, 1994), a situation that explains why over 100 alternative methods have been developed these last three decades, and most of the algorithms currently utilized are heuristic and combinatorial optimization algorithms for obtaining the approximation of optimal solution, imposed by the NP-complete nature of the problem. (Chatzou *et al.*, 2016)

We developed a new method that performs pairwise alignment in O(*n*) time between two highly similar sequences based on a suffix tree structure, which is well-known for its string processing power. Here, we take two sequences which have at least 56.8 percent, which is calculated in the experiment sector, nucleotides in common as highly similar sequences. We also designed and implemented a novel MSA method called MASC (Multiple sequence Alignment based on a Suffix tree and Center-star strategy), which has extremely high performance and accuracy with our new pairwise alignment algorithm. Furthermore, we implemented MASC on the Spark parallel distributed framework for use with massive scale sequence data and accelerate the method by using multiple computing node parallel programming.

1. **Related Work**

Multiple Sequence Alignment to be one of the most widely used modeling methods in biology, which is indeed an important modeling tool whose development has required addressing a very complex combination of computational and biological problems. (Chatzou *et al.*, 2016)

The most commonly used heuristic methods involve a progressive-alignment strategy (Hogeweg and Hesper, 1984), iterative-alignment strategy (Wallace *et al.*, 2005), or center-star strategy (Zou *et al.*, 2009). ClustalW (Thompson *et al.*, 1994) is the most widely used implementation of the progressive-alignment strategy. MAFFT (Lounkine *et al.*, 2012) is quite fast, using a fast Fourier transform algorithm along with a progressive-alignment strategy.

Progressive-alignment generates a quasi-phylogenetic guide tree among the sequences and gradually builds up the alignment in a pairwise fashion, following the order provided by the tree. Although successful for a wide variety of cases, method suffers from greediness. Errors made in initial alignments cannot be rectified later as the remainder of the sequences are added (Notredame, C., Higgins, D. G., & Heringa *et al.*, 2000). Apart from this, progressive alignment is also quite time-consuming when dealing with long sequences, due to its nonlinear time complexity. (Zou *et al.*, 2012)

The iterative strategy is an interesting alternative method, which is based on tree-based progressive strategy and involves re-estimating trees and alignments until both converge. Iterative strategies do not provide any guarantee of an optimal solution, but are reasonably robust, and are much less sensitive to the number of sequences than their deterministic counterparts (Gotoh, 1996). PRRN (Gotoh, 1996) and MUSCLE (Edgar, 2004) employ an iterative-alignment strategy.

The star alignment strategy, which is utilized in our project, is a fast method for solving MSA, and suitable for highly similar sequences. The main approach underlying the center star method is to transform MSA into pairwise alignment based on a ‘center sequence’. The center sequence is selected, and other sequences are pairwise aligned to the center sequence. Then, all of the inserted spaces are summed to obtain the final MSA result. (Zou *et al.*, 2015) The primary time cost in the star alignment strategy occurs during the construction of the similarity matrix, an upper triangular matrix that stores the similarity scores of each of the two sequences. The similarity score is calculated by pairwise alignment. This matrix is used to discover the center sequence. Assume that it takes O(*t*) to run a pairwise alignment between two sequences. The complexity of the algorithm is O() where m is the number of sequences. Therefore, running MSA on a set of homologous sequences using the star alignment strategy reduces the time cost to O(*mt*), which is much faster than both the progressive and iterative strategies.准备删除

Regardless of which heuristic method used, the main simplification idea common to all is to reduce the MSA to a series of pairwise sequence alignments. Consequently, pairwise alignment is a dominant component of all MSA techniques. Traditional dynamic programming algorithms, such as Smith–Waterman (Smith and Waterman, 1981) and Needleman–Wunsch (Needleman and Wunsch, 1970), require O() temporal and spatial complexity to perform pairwise alignment, where *n* is the maximum length of two sequences. Other, faster algorithms have been developed, such as MAFFT (Lounkine *et al.*, 2012), which is O(*n*log*n*). These algorithms work extremely well on conventional tasks with multiple single protein sequences, cDNA sequences, or relatively short genomic DNA sequences containing a single gene and simple intron interruptions, but in most cases are ineffective for aligning very long sequences.[为什么MAFFT在长序列的情况下fail。这个是需要给出说明的也就是要给出相应的理论证明或者实验上的证明] Furthermore, when performing the pairwise alignment step between two very long sequences, such as mtDNA (Iborra *et al.*, 2004), or whole genomes, many algorithms either run out of memory or take too long to complete (Delcher *et al.*, 1999). To the best of our knowledge, there is no efficient and effective pairwise alignment method that requires O(*n*) time complexity.

1. **Algorithms**

MASC uses a combination of three ideas: First, a pairwise alignment algorithm was developed, with time complexity O(*n*), based on powerful suffix tree algorithm. Then our new, center-star strategy algorithm is used for MSA, which decreases the time cost to O(*mn*), where *m* is the number of sequences and *n* is the average length of the sequences. Finally, MASC is implemented on the Spark distributed parallel framework.

*3.1. Suffix Tree Pairwise Alignment*

A suffix tree is a compressed trie (Aho and Corasick, 1975) containing all the suffixes of a given text as keys, and positions in the text as values. Suffix trees allow particularly fast implementations of many important string operations (Baeza-Yates and Gonnet, 1996). The suffix tree for the string S of length *n* is defined as follows:

|  |
| --- |
| Definition:  The tree has exactly *n* leaves numbered from 1 to *n*. Except for the root, every internal node has at least two children. Each edge is labeled with a non-empty substring of S. No two edges starting out of a node can have string-labels beginning with the same character. The string obtained by concatenating all the string-labels found on the path from the root to leaf *i* spells out suffix S[*i*..*n*], a substring also a suffix of S starts from ith character in S to the end of S, for *i* from 1 to *n*. |

Ukkonen developed a new algorithm that can construct a suffix tree in linear time with O(*n*) computational space from a string S, where *n* is the length of S. Ukkonen’s algorithm (Ukkonen, 1995) also reduces suffix tree construction to O(*n*) time, for constant-size alphabets, and O(*n*log*n*) in general. Within a bioinformatics context, the tree alphabets consist of {A, G, C, T} or {A, G, C, U} for DNA or RNA, respectively, or the 20 amino acid symbols for proteins. Therefore, the time cost is linear for all molecular biological sequences (Barsky *et al.*, 2008).

For a string's suffix tree contains all its suffixes, each leaf node represents a unique suffix. A suffix is simply a subsequence that begins at any position in the sequence and extend to the end of the sequence. For a pattern, for which we want to search, can be taken as a prefix of a suffix. A prefix is simply a subsequence that begins at the beginning of the sequence and extend to any position of the sequence. A pattern can be found in linear time by traversing a unique path in the tree from the root node to a inner node or a leaf node. In our work, suffix tree is used to find out the common substrings of two biological sequences in a pairwise alignment.

As with running a pairwise alignment of sequence A and sequence B, it is assumed that the sequences to be compared are highly similar. Therefore, there are many common substrings in both A and B. The alignment process consists of the following steps, which are shown in Figure 1.



**Figure 1** Pairwise alignment process using a suffix tree

The first step of our method is to build a suffix tree from one sequence. Here S1 is assumed as the chosen one, and the tree’s name is tree-S1. The tree construction consumes O(n) time with Ukkonen’s algorithm, n is the length of S1. Then we use a process to pick out common strings. The pseudocode is shown below:

|  |
| --- |
| Input: S1，S2: two Strings;tree-S1: the suffix tree of S1;  Output: result\_list: a list of substrings' information, whose element is composed of (a substring's start position in S1, a substring's start position in S2, the length of substring)  index=0;  while(index<S2's length){  /\*select\_prefix is a function to find the S2's longest common prefix with S2 using suffix tree tree-S1. If a prefix is found the function will return the start postion of the substring in S1 and the length of the prefix，otherwise a tuple of [-1,0] will be returned\*/  st,len=tree-S1.select\_prefix(S2,index);  /\*A common prefix(substring) is found\*/  if(st!=-1){  Record the start postion of the prefix in both S1(st) and S2(index) in result-list;  Record the length of the prefix in result-list;  index+=len;  }  /\*Common prefix is not found\*/  else{  index++;  }  return result-list;  } |

A function named select\_prefix, which is a member function of suffix tree data structure, is applied to find the longest common prefix between a given input string and a suffix in the tree. When we search common substrings between a string and a suffix tree, the select\_prefix function is used to find a common prefix, then the common prefix is skipped and reused select\_prefix to find another common prefix. The process would be repeated until all character in S2 in scanned. Thus, in a single scan of string S2, all common substrings can be identified.这里要用一个图来举例说明。

The process of picking out common strings in S1 and S2 with sequence S2 and tree-S1 is run to form two common substring sets, set sub-S1 and set sub-S2. The next step drops out ineligible substrings in sub-S1 and sub-S2. Our standard for what constitutes eligible common substrings is described below.

The first standard is that common substrings cannot be too short. We set the length threshold at 20bp[这个数值是怎么来的？？？这个还是问邹老师吧] for nucleotide sequences. This is because very short strings are extremely common in nucleotide sequences. Very short common substrings between sub-S1 and sub-S2 would not correspond with one-to-one homology, which would lead to misalignment.

The second standard is that the starting positions of two matching substrings cannot be too far away. The absolute value of the difference between the starting positions of the two matching substrings must be less than their length. For highly similar sequences, most part of the sequences are the same. The matching of remote substrings would enlarge the areas which need to be aligned by dynamic programming algorithm so that the time of alignment is extended.

After finding all common substrings and discard the substrings that do not meet the first two criteria. These substrings are sorted according to their position in S1. In some cases, the order of matching substrings in S2 may be reversed. The position of substrings are not in increasing order. This situation is diagramed below in Figure 2. Although this situation does not occur often in highly similar sequences, we still have to deal with it. The LIS algorithm (DanGusfield, 1997) is employed to find the largest set of substrings which are in ascending order in both S1 and S2. As shown in Figure 2.[加入样例的图,下面这张图要修改，加入更多的内容，比如列序，提取]



**Figure 2** Two reversed matched pairs

Next, the process of aligning the remaining unmatched substrings between S1 and S2 is run. Finally, all substrings are concatenated to form the aligned sequences.

The construction of the suffix tree can be accomplished in O(*n*) time where *n* is the length of the sequence. Picking all common substrings can be done in a single scan of string S2 which can be accomplished in O(*n*) time where n is the length of S2. The alignments among unmatched substrings can be finished in O(), where *k* is the number of substrings and *m* is the maximal length. For highly similar sequences, is far less than *n*, so the total time complexity is O(*n*).

*3.2. Center-Star Strategy*

The center-star strategy (Zou *et al.*, 2009) is a method that can solve the MSA problem with extremely high performance. And, the strategy’s performance improves with increasingly similar sequences. For a set of sequences S = {,,…,,}, the center sequence Scenter in S needs to be found, which can satisfy the following formula:

The formula would help find out the sequence which is similar to all others. Here is the similarity of two sequences si and sj. Pairwise alignments between any two sequence in S are carried out, and a similar matrix, which is an upper triangular matrix that stores the similarity scores of each of the two sequences in set S, is calculated first. Then the matrix is applied to compute the sum of scores between one sequence and others. The sequence which has minimum sum of scores is selected as the center sequence.

After that, pairwise alignments are carried out between the center and the other sequences, one by one, Each pairwise alignment gives two consequent sequences, one is the result of center sequence after aligning(inserting the gap), the other one is another sequence after aligning. The consequent sequences are collected into two sets, centerS and S’. centerS is the set of center sequences after pairwise alignments, and the S' is the set of consequent sequences of all other sequences. Next, all the inserted spaces in the sequence of centerS are summed to obtain a result sequence of center named center’. At last, pairwise alignments between center’ and every sequence in S’ is carried out to get the final multiple sequence alignment result. The center-star strategy is shown in following algorithm outline:

Input: sequence set S= {, , …, }

Output: aligned sequence set S’’= {, , …, }

1. Run pairwise alignment between every two sequences, and construct similarity matrix W, which is an upper triangular matrix that stores the similarity scores of each of the two sequences in set S.
2. Select using W.
3. Run pairwise alignments between the center and the other sequences to get set centerS, which is the set of the aligned center sequence, and S′, which is the set of other aligned sequences.
4. Sum all the inserted spaces to get the aligned center named .
5. Run pairwise alignment between and the other sequences again to get the result of the MSA.
6. Output the result.

When performing MSA among a set of similar sequences, it takes O(*n*) to finish a pairwise alignment with our suffix tree method previously illustrated. The construction of similarity matrix W can be accomplished in O(), where *m* is a cardinal number of sequence set S. Pairwise alignments among the center and the other sequences can be finished in O(*mn*) time. The total time cost is no more than O(). The process of finding the center sequence takes most of the time. When aligning similar sequences, every sequence can be regarded as an average one. This means that each sequence can be selected as a center, so that selecting a random one is a good choice, which reduces the time complexity to O(*mn*).

*3.3 Distributed and Parallel Implementation with Spark*

As MASC can perform MSA in linear time with highly similar sequences, it has enormous potential for very-high-throughput cases. We can align a larger scale of extraordinarily long sequences with better performance than other existing algorithms. However, in reality an ordinary computer cannot keep that many long sequences in memory, possibly leading to a memory crash, or wasting excessive time shifting between physical and virtual memory. Consequentially, we needed to solve the memory storage problem, and parallelize our method using the power of a distributed parallel framework.

There are many frameworks available for handling large scale data. MapReduce (Dean and Ghemawat, 2004) and Spark (Zaharia *et al.*, 2010) are the most powerful and popular. Spark is more suitable for this work, due to its memory computation characteristics.

Spark is a MapReduce type framework that can carry on all computation in memory, without needing to save intermediate results. The main abstraction Spark provides is a resilient distributed dataset (RDD), which is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. RDDs are created by starting with a file in the file system, or an existing collection in the driver program, and transforming it.

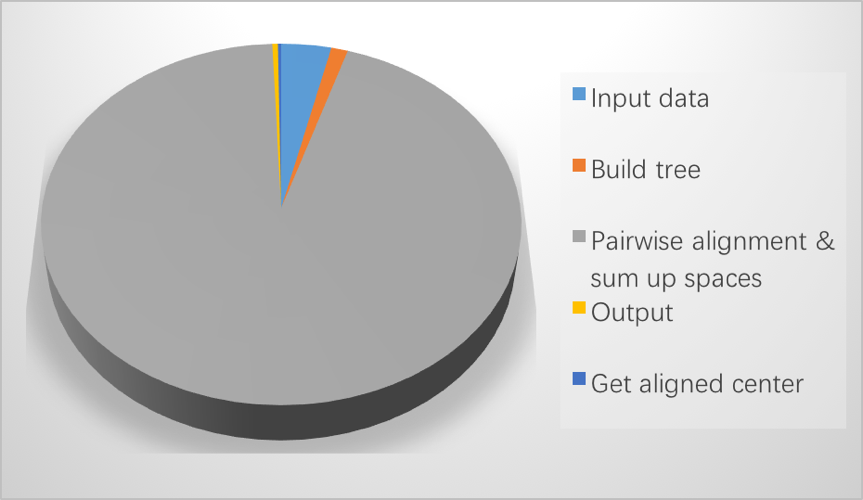
The serial version of our method is illustrated in Figure 3.



**Figure 3** MASC flowchart

After inputting the data, the first task is to select a random sequence as the center and build the suffix tree. Then, the pairwise alignments between the center and all other sequences are run. Next, the process of summing up the spaces among sequences is accomplished. Finally, the result is output.

We analyzed data flow to find ‘hot spots’ within the program, the processes of “pairwise alignment and sum up spaces”, which occupies most of the computational time, as shown in Figure 4.



**Figure 4** Time cost of MASC steps

Fortunately, in MASC the process of alignment between the center and the other sequences is readily parallelized, because there is no data dependency. Concurrently the remaining serial processes do not occupy too much time and memory. [这句话要重新，不知道为什么？？语法错误？？]

According to Amdahl’s law (Moncrieff *et al.*, 1996) :

In the formula above, T is the total time that a program consumes, which consists of and . is the time consumed by the part of program that can be parallelized. is the time consumed by the part of program that must be run in serial. And p is the number of processors. So the speedup of our method is:

So the limitation of the speedup of our program is 16 in theory, which means the parallelization of MASC seemed very promising.[被指出没有用数学证明，在这里加入amdal定理。还有一个看不懂的问题。。另外figure 5 中有拼写错误]



**Figure 5** Implementation of MASC on Spark

MASC was therefore developed within the Spark framework, and the process is shown in Figure 5. Our Spark center-star MSA has two stages. Initially, data is input from the local file system or an Apache Hadoop Distributed File System (HDFS), and a format examination is performed. Next, the sequence array is parallelized, which consists of converting the string lists to sequence string RDDs. Concurrently, the program needs to choose a center sequence, which, as previously explained, is randomly chosen, because any sequence can be regarded as average, if all are sufficiently similar. This randomly selected sequence is then used to construct a center sequence suffix tree, and pairwise alignments are run between that suffix tree and all other sequences, using Spark, making use of the algorithm previously illustrated. In this step, the suffix tree is used to get matching substrings in each sequence serially, because this step runs extremely fast, and wasting memory can be avoided by keeping a single copy of the center suffix tree. When the information regarding all the matched substrings is obtained, a parallel process of aligning the unmatched substrings is run by passing the function to the Spark transformation, which implements the Needleman–Wunsch algorithm. In the previous step, the sequence string RDDs are transformed into pair RDDs of spaces in the center sequence and spaces in all other sequences. Next, the RDDs of spaces in the center sequence are collected to make the center aligned. Then the aligned center sequence is broadcast to all executors, and align all other sequences. In this step the pair RDDs of spaces are transformed into string RDDs of aligned sequences, and the results are stored. Finally, the aligned string RDDs are collected and are output to the local file system. The data flow and operations are shown in Figure 6.



**Figure 6** Data flow and operations

In Figure 6, all the nodes represent data and the edges represent the operations. The cycle elements are the RDDs in Spark that are distributed in the executors, the rectangle elements are datasets in the driver’s memory. The operations signed by solid lines represent the transformations in Spark that transfer RDDs to subsequent RDDs and the dotted line edges represent the action operations that convert RDDs into dataset in driver memory.

1. **Experiments**

*4.1. Datasets and Measurements*

Balibase (Thompson *et al.*, 2005) is regarded as the golden benchmark for most MSA research. However, the database is relatively small, and is only suitable for protein sequence alignment. Because there is no benchmark dataset for addressing the large-scale nucleotide MSA problem, we employ human mitochondrial genomes (mt genomes) and 16S rRNAs as the test data in our experiment. Human mt genome MSA analysis is necessary for detecting mtSNP sites, which are associated with Alzheimer’s Disease, Parkinson’s Disease, and Type 2 Diabetes (Tanaka *et al.*, 2004).[这个地方的问题我不懂，要问老师] Our human mt genome dataset contains 672 highly similar mt genome sequences, with a maximal length of 16,579 bp, and a minimal length of 16,556 bp. With the aim of testing the performance of our program with large-scale data, we duplicate the mt genomes 20 times, 50 times, and 100 times to enlarge the test set.

Table 1 Detailed information on the experimental DNA dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Max length | Min length | Average length | Sequence number | File size |
| mt genome(1x) | 16579 | 16556 | 16569.7 | 672 | 10MB |
| mt genome(20x) | 13440 | 213MB |
| mt genome(50x) | 33600 | 532MB |
| mt genome(100x) | 67200 | 1.1GB |

Our main project purpose is to accelerate the MSA process, and improve the capacity to handle massive data. Therefore, we focus our attention on running time and throughput. We choose the sum-of-pairs (SP) value (Zou *et al.*, 2009), which is a measurement of the quality of an MSA, for measuring alignment performance. The SP value is an integer representing the sum of every pairwise alignment score from an MSA. For our purposes, in a pairwise nucleotide sequence alignment, if two nucleotides from the same column are different, one is added to the SP value, while if a space is inserted, two is added to the score,[这里的问题是，gap为什么比mismatch得分高？？] otherwise, if the two nucleotides are the same, the score remains unchanged. Thus, the SP value will be a positive integer, and the lower the SP value, the better the quality of the MSA. However, SP values are not suited for massive MSAs because the score may become too large and exceeds the computer’s limitations. Therefore, we employ the average SP value instead, which is the SP value divided by the number of sequences (Zou et al., 2015).

*4.2. Sequence similarity analysis*

MASC is a method to do multiple sequence alignment among highly similar sequences. In order to find out what kind of sequences that are suitable to be aligned by MASC, an experiment is designed to quantify the similarity of sequences. A 1000bp long sequence is taken as basic sequence in the experiment. Then variations are carried on the copies of basic sequence. The dataset is formed by basic sequence and modified copies. The variation is to change one residue of basic sequence to an element of {‘A’, ‘G’, ‘C’, ‘T’, ‘-‘} where ‘-‘ means to delete the residue and concatenate to substrings. The quotient of the number of residues which are changed divided by length of basic sequence is called mutation rate. In the experiment, mutation rate ranges from 0 to 100%. The result shows that MASC works when the mutation rate is no more that 54% and the output is better that MAFFT measure by average SP value (Zou *et al.*, 2009). The result of comparison is shown in Figure7.



**Figure 7** SP value of MASC and MAFFT with mutation rate changed

[缺少轴上的单位，还缺一个算法的结果，需要补充一下]

The less SP value is on behalf of better accuracy. MASC is more accurate than MAFFT when the mutation rate is no more than 54%. Due to the variation is selected from {A, G, C, T, -} in random, five elements have the same probability to be chosen, so that the residue may be unchanged. It means that the MASC works when there are more than 56.8% residues are same between sequences that need to be aligned.

*4.3. Comparison with State-of-Art Tools*

Most of the available state-of-the-art MSA software tools cannot address large-scale data. Therefore, we only perform comparisons with MAFFT and KAlign. KAlign and MAFFT are run on single nodes, without any parallel operations. Therefore, for fairness, we carry out all the experiments on the same server (Intel® Xeon® [CPUE7-8890v3 @ 2.5 GHz](mailto:CPUE7-8890v3@2.5GHz) with 2 TB total memory, and the Red Hat Enterprise Linux Server release 7.1 operating system). Table 2 shows time consumption for the various human mitochondrial genome datasets.

Table 2 Time consumption for different MSA tools using human mitochondrial genome datasets of different multiplicities

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 672(1x) | 13440 (20x) | 33600 (50x) | 67200 (100x) |
| MASC-serial | 35s | 10m54s | 26m14s | 51m51s |
| MASC-spark | 7s | 1m50s | 5m11s | 8m44s |
| MAFFT | 1m59s | 3h52m14s | 21h54m18s | 3d12h41m42s |
| KAlign | 1h27m10s | --- | --- | --- |

MSA is run with different kinds of software on different datasets which contains different number of sequences. From Table 2 we can see that MAFFT and KAlign take an extremely long time to finish the MSA among long sequences, even with relatively small files. Furthermore, KAlign cannot even handle files larger than 100 MB. However, MASC (serial version) is compatible with massive files of long sequences, and the parallelized version runs extremely faster than all other programs. The parallelization analysis will be shown below.

Table 3 Accuracy of different methods

|  |  |  |  |
| --- | --- | --- | --- |
|  | MASC | MAFFT | KAlign |
| 10M(1x) | 14276 | 15202 | 14809 |

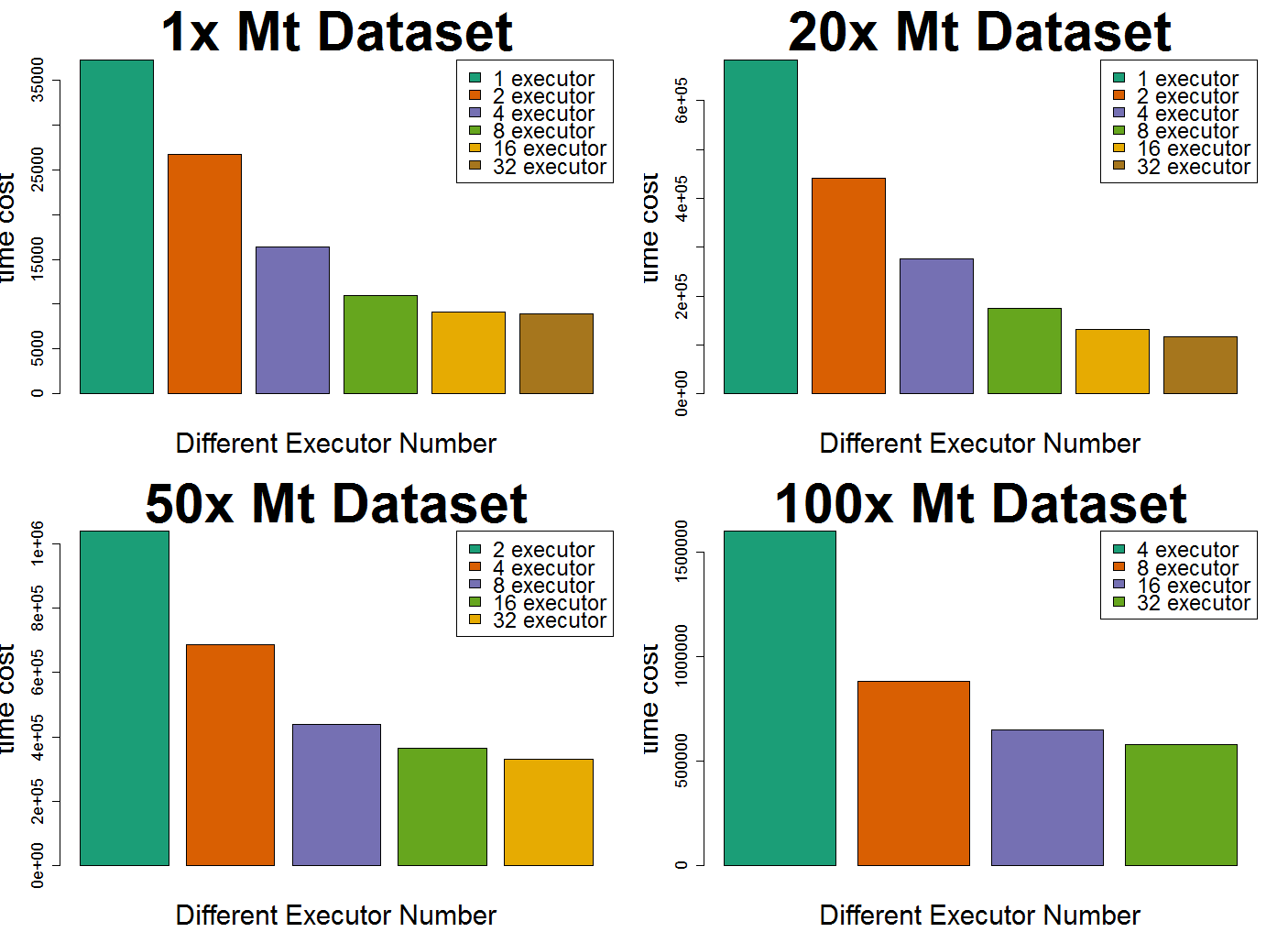
Table 3 shows a comparison of the average SP values among the different programs. As we have described previously, a lower SP value means better accuracy. Because the human mitochondrial genome dataset sequences are highly similar, the different programs perform similarly. However, the data in Table 3 shows that our method is somewhat more accurate than the other two methods with this dataset.

MASC can be used for any file, no matter how large it is. In comparison, MAFFT cannot be used for files larger than 1 GB, and KAlign cannot be used for files larger than 10MB. MASC is clearly superior for processing massive MSAs.

*4.4. Speedup of Spark*

We demonstrate the speedup and scalability of the Spark version of MASC in this section. We perform our experiments on a server with 16 CPUs and 2 TB of memory (same specifications as previous).

We test the 1×, 20×, 50×, and 100× human mt genome datasets with different modes that run various numbers of executors. Figure 7 shows the results of those experiments. Each combination of executors number(Zaharia *et al.*, 2010) and dataset was run over different 50 times. Here executors mean threads that do pairwise alignments.



**Figure 8** Running times for different mt genome datasets with different modes

In Figure 8, each bar plot shows the performance of MASC on different datasets. Each column shows the average time cost of the program running in different environments (which can be regarded as different executors on the cluster). The y-axis unit is seconds. We do not test the one and two executor modes for the 100× dataset, and the one executor mode for the 50× dataset, because it is too difficult for one node keeping such a huge copy of massive data in memory. We see a remarkable speedup by Spark parallelization when we use no more than 32 executors at the same time. The speedup can be over ten sometimes, with an average total value of eight. The scalability is well-maintained, and efficiency remains constant, while we enlarge the dataset and the number of executors at the same time.

1. **Conclusion**

Multiple sequence alignment is an important and fundamental bioinformatics tool. We propose MASC in this study, which can perform MSA in O(*mn*) time among highly similar sequences, where *m* is the number of sequences in the dataset, and *n* is the average length of the sequences. MASC has very high accuracy and performance. The core idea of our method is to accelerate the MSA process using three steps:

We speed up the process of pairwise alignment based on a suffix tree, which is a powerful data structure for handling strings. Time complexity is O(*n*) at the most, when aligning pairs of similar sequences based on suffix trees. A center-star strategy is then employed as a heuristic to reduce the MSA problem to pairwise alignments. MASC can be accomplished in O(*mn*) time when the sequences are highly similar. There is no loss in accuracy. Along with MASC’s extremely high performance, we used the distributed parallel computing framework Spark to enlarge the memory of the system, and, in this way, the throughput of our program is substantially increased.

Extensive experiments with MASC were then performed. First, the accuracy and performance of our method was tested compared with other state-of-art tools. The scope of the comparison was quite limited, because most available tools are not optimized for performance nor efficiency. MAFFT and KAlign are two optimized tools available that were selected for comparison. The results of our experiments show that we have made great progress, and that our method has better accuracy than the other two methods with our sample datasets, as indicated by lower average SP values.

MASC has been implemented on Spark and HDFS to handle the increasingly expansive data in the field of biology and bioinformatics. The method is very suitable for parallelization because the pairwise alignments between sequences are independent and are reduced from MSA by our center-star strategy. In practice, the Spark version of MASC has great speedup and scalability. Both the single thread and the parallel tools are coded with Java, which works on multiple operating systems. Java 1.8 and Spark 2.0 are prerequisites for its operation. The codes and tools are accessible free of charge at <https://github.com/suwenhecn/MASC> .

Conclusion中要加入局限性和之后工作的讨论

**Acknowledgement**  
The work was supported by the Natural Science Foundation of China (No. 61370010, 61272056, U1435222, 61133005) and Guangdong Provincial Department of Science and Technology under grant No. 2016B090918122. Thanks to Ms Dong’s help and her companion.

**Reference**

Aho,A. V. and Corasick,M.J. (1975) Efficient string matching: an aid to bibliographic search. *Commun. ACM*, **18**, 333–340.

Baeza-Yates,R. a. and Gonnet,G.H. (1996) Fast text searching for regular expressions or automaton searching on tries. *J. ACM*, **43**, 915–936.

Barsky,M. *et al.* (2008) A new method for indexing genomes using on-disk suffix trees. *Proceeding 17th ACM Conf. Inf. Knowl. Manag.*, 649–658.

Chatzou,M. *et al.* (2016) Multiple sequence alignment modeling: methods and applications. *Brief. Bioinform.*, **17**, 1009–1023.

DanGusfield (1997) Algorithms on strings, trees, and sequences Cambridge University Press.

Dean,J. and Ghemawat,S. (2004) MapReduce: Simplified Data Processing on Large Clusters. *Proc. OSDI - Symp. Oper. Syst. Des. Implement.*, 137–149.

Delcher,A.L. *et al.* (1999) Alignment of whole genomes. *Nucleic Acids Res.*, **27**, 2369–2376.

Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Gotoh,O. (1996) Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments. *J. Mol. Biol.*, **264**, 823–838.

Hogeweg,P. and Hesper,B. (1984) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *J. Mol. Evol.*, **20**, 175–186.

Iborra,F.J. *et al.* (2004) The functional organization of mitochondrial genomes in human cells. *BMC Biol.*, **2**, 1–14.

Lounkine,E. *et al.* (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, **486**, 361–7.

Moncrieff,D. *et al.* (1996) Heterogeneous computing machines and Amdahl’s law. *Parallel Comput.*, **22**, 407–413.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similiarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Notredame, C., Higgins, D. G., & Heringa,J. *et al.* (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Tanaka,M. *et al.* (2004) Mitochondrial Genome Variation in Eastern Asia and the Peopling of Japan Mitochondrial Genome Variation in Eastern Asia and the Peopling of Japan. *Genome Res.*, 1832–1850.

Taylor,W.R. (1990) Hierarchical method to align large numbers of biological sequences. *Methods Enzymol.*, **183**, 456–474.

Thompson,J.D. *et al.* (2005) BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins Struct. Funct. Genet.*, **61**, 127–136.

Thompson,J.D. *et al.* (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Ukkonen,E. (1995) On-line construction of suffix trees. *Algorithmica*, **14**, 249–260.

Wallace,I.M. *et al.* (2005) Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics*, **21**, 1408–1414.

Wang,L. and Jiang,T. (1994) On the complexity of multiple sequence alignment. *J Comput Biol*, **1**, 337–348.

Zaharia,M. *et al.* (2010) Spark : Cluster Computing with Working Sets. *HotCloud’10 Proc. 2nd USENIX Conf. Hot Top. cloud Comput.*, 10.

Zou,Q. *et al.* (2012) A Novel Center Star Multiple Sequence Alignment Algorithm Based on Affine Gap Penalty and K-Band. *Phys. Procedia*, **33**, 322–327.

Zou,Q. *et al.* (2009) An Algorithm for DNA Multiple Sequence Alignment Based on Center Star Method and Keyword Tree. *Acta Electron. Sin.*, **38**, 1746–1750.

Zou,Q. *et al.* (2015) HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. *Bioinformatics*, **31**, 2475–2481.