

Response to Reviewer's Comments

TITLE : Alignment-free Sequence Searching over Whole Genomes Using 3D Random plot of Query DNA Sequences

Dear Reviewers,

We greatly appreciate your reviewing our manuscript and providing constructive comments and suggestions. All of your comments are helpful for improving the quality of the manuscript.

In this revised draft, we added some detailed descriptions in order to clarify several important contributions especially focusing on the similarity sequence search model part. Also we corrected some grammatical errors and typos that you pointed out.

The detailed answers to your comments are listed in the following table:

Reviewer B

Comments	Answer
1. The default grid size is 500x500x500. How does it affect the sequence comparison?	<p>As you pointed out, there is a trade-off regarding the grid size. With a larger-size grid, the sequences are represented more compactly. This makes easier for users to find the characteristics of the sequences. Also, the similarity computation can be performed more efficiently in terms of time and space. On the other hands, a large-grid representation loses the detailed features of the sequences so the computed similarity score can be inaccurate.</p> <p>The default grid size 500x500x500 is what we empirically figured out at which this trade off is well balanced for the sequences used in the experiments.</p> <p>To clarify this part, we added a description in the draft as follows:</p> <p>In this way, the transformed random plot is visualized in an appropriate sized three-dimensional grid. The default grid size 500x500x500 is what we empirically figured out at which this trade off between speed and correctness of comparison is well balanced for the sequences used in the experiments.</p>

<p>2. Some techniques are based on [14]. So what's the major difference of this paper and [14]?</p>	<p>The 2-mer vector allocation model used in [14] was also used in this paper. However, the similarity comparison method of [14] uses the ratio of the area of the space common to the two plots to be compared, so it compares the sequence as a whole.</p> <p>Therefore, even if the lengths of two sequences are different and the compression ratios are visualized differently, the exact similarity can not be evaluated because the area of the plot originating from the origin is compared. In addition, there is a similar section in the area where the plot area does not overlap. [14] does not consider this problem, and therefore there is a limit in that the similarity calculation does not reflect reality.</p> <p>In this paper, we divide the plot into unit sizes and compare the vector components of each plot. Therefore, it is possible to calculate partly similar components because the direction and degree of the plot are considered regardless of overlapping areas. In addition, since the compression rate of the reference sequence is the same based on the query sequence, the approximate position of the actual similar region can be searched in b.p. units regardless of the length difference of the two sequences.</p> <p>To clarify this part, we added a description in the draft as follows:</p> <p>Compared to [14], we present an improved similarity computation algorithm that considers input sequences with different lengths. We show the effectiveness of the proposed method with experiments on searching for short query sequences on a long sequence.</p>
<p>3. In Definition 2, the distance is defined recursively. What's the stopping criterion for the recursion? From Algorithm 1, it seems this is determined by the length of vector, but not clarified in the paper.</p>	<p>If the length of divided vector drops below the appropriate length D, the recursion is aborted. In this case, the threshold D value is set to 100 times the unit size, where unit size is the number of bp per pixel when visualized. That is, the unit size vector (100px) is compared through recursion, and the size of bp included in the unit size vector differs depending on the compression ratio.</p> <p>To clarify this part, we added a description in the draft as follows:</p> <p>If the length of divided vector drops below the appropriate</p>

	length D, the recursion is aborted. In this paper, the threshold D value is set to 100 times the unit size, where unit size is the number of bp per pixel when visualized. The D value was determined experimentally because at least the length of the vector was more than 100px, meaningful comparison was possible.
<p>4. On page 8, what do you mean by "expression 3.5"? Format and grammatical errors are rampant in this manuscript. Please consider a careful proofreading.</p> <p>1. Title: "random plot" -> "Random Plot".</p> <p>...</p>	<p>Thank you for finding typos. We corrected the 14 points that you pointed out, and we also corrected all parts of the paper that have the same problem.</p>

Reviewer C:

Comments	Answer
<p>1. The author should unify if they have the same meanings:</p> <p>walk-plot, random plot and random walk plot</p>	<p>We unified those words in 'random plot'.</p>
<p>2. It was not clear that what 3 dimensions are? and how to relate them to the original sequence?</p>	<p>We transform a sequence of biological symbols into 3-dimensional vectors in order with the rules described in Table 1. The base vectors in the transformation rule are determined empirically in order to discriminate different sequences effectively. As a result, similar sequences are likely to produce similar walk plots.</p> <p>For example, if a random plot tends to proceed in the positive direction of the X axis, this means that the sequence contains a lot of 'A'bases. Also, visualization of two genome sequences suggests that the sequence of the random plot is similar,</p>

	<p>indicating that the bases in the actual sequence are similar.</p> <p>To clarify this part, we added a description in the draft as follows:</p> <p>This vector transformation rule are determined empirically in order to discriminate different sequences effectively. As Figure 6, similar sequences are likely to produce similar walk plots.</p>
<p>3. Why was 500x500x500 grid mentioned in the manuscript for short sequences? What are for large sequences?</p>	<p>As you pointed out, there is a trade-off regarding the grid size with difference size sequence process. The default grid size 500x500x500 is what we empirically figured out at which this trade off is well balanced for the sequences used in the experiments.</p> <p>To clarify this part, we added a description in the draft as follows:</p> <p>In this way, the transformed random plot is visualized in an appropriate sized three-dimensional grid. The default grid size 500x500x500 is what we empirically figured out at which this trade off between speed and correctness of comparison is well balanced for the sequences used in the experiments.</p>
<p>4. The meaning of "Sim" measure was not clear? How did the author calculate similarity between a geometric-based and the original sequence?</p>	<p>"Sim" is the result of similarity described in 3.4.</p> <p>To clarify this part, we changed the function name COMPARE into SIM in Algorithm 1.</p> <p>And we added a description in the draft as follows:</p> <p>This modification rate is expressed as 'M' (M.Rate) in Table 3 and 4. 'M' (M.Rate) refers to the modified ratio of the number of B.P. on origin sequence. For verification of the similarity comparison model, this rate was set higher gradually as the experiment was repeated.</p>
<p>5. More importantly, the author did not show any comparison with other state-of-the-art methods in this important research field.</p>	<p>To best of our knowledge, this is the first work that considers both visualization and computational similarity comparison on large sequences. As a result, we give a qualitative comparison with other previous work related to each aspect of our contribution as in Table 1 which we added in this revision.</p> <p>For the visualization aspect, our method uses a sufficient</p>

number of base vectors to represent the sequence in 3-dimensional space. It enables to visualize both global and local features of the input sequences compared to other plot-based visualization methods. For the similarity computation aspect, most of methods are based on the sequence alignment algorithm that requires a quadratic time complexity, which is infeasible for comparing large sequences.

Table 1: Functional Performance of Previous Research

Research	Plotting space dimension	Supports large-scale sequence	Global similarity compute	Local similarity compute
BLAST [1]	N/A	\triangle	O	O
Compact 2D [4]	2D	O	O	X
H-L Curve [7]	2D	\triangle	X	X
Bo Liao [17]	3D	\triangle	O	X
3D Random [14]	3D	O	O	X
Proposed	3D	O	O	O