

MISHIMA: a Method for Identifying Sequence History In terms of Multiple Alignment

Kirill Kryukov¹, Naruya Saitou²

Keywords: multiple sequence alignment, divide and conquer, dictionary of motifs

1 Introduction.

Most of multiple sequence alignment methods in current use depend on a pairwise sequence comparison and/or progressive alignment technique. This leads to large computational complexity in case of large number of sequences and/or very long sequences. We present a new method of multiple sequence alignment, that does not depend on pairwise sequence comparison. Instead a heuristic method is used to quickly find homology shared by multiple sequences. Divide and conquer approach is then applied to divide the sequences into fragments that can be aligned independently by an external alignment program. After that partial alignments are assembled together to form a complete alignment of the original sequences.

2 Dictionary-based alignment method.

The core of MISHIMA method is its way of locating homology: At first, dictionary of all sequence motifs of length up to K nucleotides is constructed. The dictionary keeps some simple information about each motif: how many times a motif was found in sequence data, how many sequences contain this motif, etc. This information can be collected with only one reading pass through the sequence data. After that the dictionary is searched for motifs that are rarely occurring on the sequence data, yet occur in several sequences. Such motifs are likely to represent a homology signal. After those 'good' motifs are found, input sequences are read again to find where exactly the 'good' motifs are located in sequence data. Based on the coordinate information a selection of 'seeds', motifs representing most probable homology, is extracted.

Dictionary of motifs is, of course, a memory-expensive data structure, its memory requirement is $O(4^K)$, where K is the motif length. Yet still dictionary analysis with $K=12$ is possible on a decent desktop computer with 1 GB of RAM. The dictionary is only used for the first step of operation: once a set of 'good' motifs is selected, a dictionary is not needed anymore and can be disposed, freeing memory for further analysis steps. The important advantage of this technique is that all of its steps require linear or shorter time.

After the set of 'seeds' is found the divide-and-conquer approach to the multiple sequence alignment is applied, similarly to [1]. The sequences are divided into segments, that are aligned independently from each other, using external alignment program. ClustalW [2] was used in this study. Partial alignments are then concatenated together to construct a complete multiple alignment.

¹ National Institute of Genetics, Mishima, Yata-1111, Shizuoka 411-8540, Japan. E-mail: kkryukov@lab.nig.ac.jp

² National Institute of Genetics, Mishima, Yata-1111, Shizuoka 411-8540, Japan. E-mail: nsaitou@genes.nig.ac.jp

The initial MISHIMA algorithm was depending on exact motif matches in homology search step. In order to increase homology search sensitivity a non-exact motif matching was added. Motifs that differ by only one substitution are now counted together during the dictionary analysis. Adding this technique requires an additional dictionary processing step, but fortunately the computations is possible in $O(1)$ time.

3 Results.

This method was tested by aligning a dataset of 10 complete mitochondrion genomes of mammals, as an example of well conserved sequences. MISHIMA implementation could successfully align the dataset, taking only two minutes. In comparison, ClustalW alone takes several hours to align this dataset. This performance is the main advantage of this new heuristic approach. This speed improvement is possible because of the linear time complexity of the MISHIMA algorithm, while most of existing algorithms take longer time, usually proportional to the square of the dataset size.

More difficult test dataset was constructed of 4 complete genomes of *Streptococcus pyogenes*, each about 2 MB long. MISHIMA method could construct the alignment, dividing the dataset into 485 segments that were aligned separately and then concatenated together. Alignment took about 6 hours on a Pentium 4 2.8 GHz machine with 1 GB of RAM. The alignment have revealed that several large insertions and deletions happened in the evolution of this species. This dataset would probably take weeks to align by progressive alignment methods, like ClustalW.

References

- [2] Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673-4680.
- [1] Tonges, U., Perrey, S.W., Stoye, J., Dress, A.W.M. 1996. A General Method for Fast Multiple Sequence Alignment. *Gene* 172(1):GC33-GC41.