**Research Paper’s Discussion**

**Written By : Khaled Omar.**

**Faculty: Computer science & Information Mansoura University.**

**Dept : Bioinformatics.**

**Level: 200**

**Paper’s Topic: Multiple Sequence Alignment (MSA) & Software tools or algorithms used for.**

**Analysis type: Sequence alignment.**

**Type of Data Used: DNA.**

**Introduction**

**-In recent years, the estimation of enormous multiple sequence alignments (MSA) is a problem requiring more techniques to be faced or solved for achieving high accuracy results. Here I’ll explain briefly two types of software tools used for MSA. The tools are: PASTA and UPP. They’re suitable for constructing alignments on enormous datasets. They can produce up to million high accuracy sequences and computed trees on these alignments are too accurate. PASTA produces the most accurate tree if the input sequences are all full-length. on the other hand, UPP provides improved accuracy compared to PASTA and other methods.**

**-There’s no doubt that MSA is a complex bioinformatics task. It’s a precursor for many processes as protein structure, function prediction and genome assembly. I ‘ll focus on high accuracy methods instead of low accuracy standard methods. In 2009, Simultaneous Alignment and Tree Estimation( SATé ) was launched to enable the co-estimation of alignments and trees on large datasets. It uses two methods: divide-and-conquer and iteration to produce high accuracy alignments on large datasets.**

**-SATé-II was launched in 2012 to improve the scalability and accuracy of SATé, but it uses up to 50,000 sequences only in spite of running on larger datasets than SATé. finally in 2014, PASTA was developed to enable analysis of larger datasets with higher accuracy.in 2015, it was discovered that SATé-II was unable to produce high accuracy alignments. Then Ultra-large** **alignments using Phylogeny-aware Profiles (UPP) is produced. It’s based on a Machine Learning (ML) technique that makes it suitable for ultra-large datasets (up to million sequences). The general divide-and-conquer strategy used in these tools is to divide the current tree, sequence into smaller subsets in each iteration. Each subset is aligned, then merged and finally a new tree is produced. SATé and PASTA differs in merging sub-alignments. they use hierarchical approach. Hierarchy reflects the tree and uses external methods as OPAL & Muscle to merge alignments. PASTA uses spanning tree to compute pairwise alignment mergers, which are combined using transcivity.**

**Here I ‘ll resume discussing briefly how MSA software tools work on DNA sequences in order to solve bioinformatics problems. I ‘d mentioned some information about PASTA and SATé tools and how they work. Now I’ll complete discussing PASTA tool. They work by iterative divide-and-conquer strategy, whose idea of working had been discussed. PASTA tool enables a selected MSA method to be run only on subsets of bounded size selected by the user. There’s PASTA GUI that has many parameters as aligner, merger, model, data type..etc.**

**Finally, it produces both alignments and trees for each iteration in a temporary files. It also outputs a config file recording all the settings used, but for large datasets, it produces Gappy alignments because it’s conservative in recognizing homologies. Now, I have finished discussing briefly PASTA tools and its output.**