**Research Paper’s Discussion**

**Written By: Khaled Omar.**

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

**Program: 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.**

**It was proved that UPP tool is more accurate than PASTA tool and suitable for large datasets, it** **builds on PASTA to improve its ability to align datasets with fragmentary sequences using Hidden Markov models (HMM), which is a probabilistic graphical model used to add sequences into multiple sequence alignment. To add a sequence S into a multiple sequence alignment A, a profile HMM is built for A, and then an optimal path through the model is found for S. Once the path is found, it defines a way of adding S into the alignment A. Note that this addition does not define an alignment between A and those letters in S that are mapped to insertion states. Thus, when using HMMs to extend A to include S, some parts of S may remain unaligned. Profile****HMMs is a collection of profile HMMs that are built using a multiple sequence alignment A, with each profile HMM in the set based on just a subset of the sequences in the set. Thus, the match states in each of the profile HMMs in the set correspond to sites in A. UPP is a combination of PASTA and an ensemble of principle HMMs, this process includes 4 steps (first 2 can be omitted if user wants to provide UPP with a pre-computed backbone alignment and tree. The process 4 steps are:**

**1.** **Given a set S of unaligned sequences, UPP starts identifying those sequences to be part of backbone alignment. This is performed by restricting S to those sequences with length within 25% of the median sequence length, and then selecting randomly sequences set from that set. Sequences number in the backbone and restrictions on what sequences can be included in the backbone can be modified by using a configuration file or input options (-B, -M, -T, and -l).**

**2.** **UPP uses PASTA to compute a MSA A and tree T on S, which are referred to as backbone alignment and backbone tree.**

**3.** **UPP builds an ensemble of profile HMMs to represent the MSA alignment A on S. It uses the tree T to break the set of sequences into disjoint subsets of bounded size, and then computes a profile HMM for each subset.**

**4.** **The remaining sequences (not in the backbone alignment) are added to A′ using the ensemble of profile HMMs (computed in Step 3), thus producing a multiple sequence alignment A on S.**

**The most important algorithmic inputs(options) in using UPP is which sequences to put in the backbone subset S′, which method to use to compute an alignment on S′, and which algorithmic parameters to use to build the ensemble of profile HMMs.** **For which sequences to put in S′, there are two decisions that need to be made: which sequences are close enough to full-length to be considered, or how many of these sequences to use for backbone alignment.**

**My Conclusion:**

**1.Bioinformatics field has vast progress and development nowadays and many new techniques and tools are released every day with their various usage.**

**2.Bioinformaticians should keep track of these new techniques and tools used for data analysis, sequence alignment, sequence assembly etc...**

**3.bioinformaticians need to learn lots of skills to be able to develop bioinformatics tools used for several purposes in bioinformatics field.**

**4.bioinformatics field is a researching field and depends on data analysis and can be used in drug design and solving most common biological problems.**

**5.there are lot of biological problems with different degree of complexity needed to be solved using bioinformatics techniques or software tools or developing new convenient ones.**

**6.every bioinformaticians should look for existing different biological problems and try to find solutions for these problems using different bioinformatics techniques, algorithms or tools or developing new tools or algorithms.**

**The End**