Technical Supplement

Efficient Algorithm Design for Pattern Discovery in Bioinformatics Sequences

Deepak Garg, Department of Computer Science and Engineering

Data Mining is becoming important due to manifold increase in data volumes. Research laboratories and biologists are inventing new sequences of various genomes. The sequence databases are increasing exponentially. The speed of increase is set to increase many fold in years to come. The existing tools are finding it difficult to process so much data with the existing algorithms. The improvements are needed for matching with the speed of increase in the data. With this intention in mind, a number of tools that are popular among the biologist community for DNA and protein sequence comparison like FASTA, RefSeq, BLAT, ClustalW, MEME and BLAST and its variants were understood in terms of their algorithms and working. Out of these tools, BLAST has been found to be more rigorous and promising. It is more robust compared to others, as some of them are tailored for specific applications. A detailed study and experimentation on BLAST has been done in this work. It includes all the variants of BLAST like BLASTN, BLASTX, BLASTP, TBLASTN, TBLASTX, WU-BLAST, PSI-BLAST, PHI-BLAST and others. The focus of experimentation and improvement has been on computational perspectives as facilities of major biological equipment are not available at this place of work.

The problem taken up for this research work is very significant in the current scenario. There has been a lot of work on the working of BLAST algorithm. The algorithm has also being improved consistently from the time of its initial development. The problem of exponential increase in the number of sequences and in the length of sequences has outpaced the research in the field of bioinformatics. To continuously find some important patterns, analogies, differences between various sequences more efforts are needed. There is a need to improve the algorithms of existing tools in general and BLAST in particular, because it is being used extensively. These were the reasons to select this problem for research work. Algorithms and statistical analyses of the search results are very important to find out many hidden information patterns that are beneficial for the biologists. After initial hiccups, many sub-problems were taken up, and after changing the algorithm, experimentation has been done. Those algorithms and statistical methods that give significant improvement compared to the results given by BLAST are included in this thesis. After discussions with various groups and people at NCBI it appeared that there are still many aspects which are still to be covered to make BLAST algorithm more efficient and effective. After detailed analysis of all parameters, it came to our mind that there are many aspects which can be improved upon to have a positive impact on the execution time or sensitivity of the resulting sequences in BLAST algorithm. It has been found after experimentation that some aspects do not improve and do not enhance the results significantly while some others have very positive impact on the performance of the algorithm.

During the work one major problem was felt that the resulting sequences were also containing many entries which are not of much use to the biologist so somehow the number of non genuine entries could be removed or reduced to the maximum extent possible. There are a number of methods which can be used for finding the areas of lesser complexity in protein sequences to rule out the non-genuine searches coming in the results of homology searching. All these algorithms use some kind of filtering or masking criteria that tends to lose information. It can result in losing some functionally useful subsequences for the biologist. The sub sequences are also excluded or included based on the statistical values that selection may not necessarily have a basis in the properties of the amino acids being compared. In this work an approach has been proposed that does not use any masking criteria and has the better performance than other algorithms available. It also has the inherent advantage of using the properties of amino acids resulting in better performance.

Another improvement in proposed method is in the dropoff score. Drop off score used to employ the scores from the standard scoring matrices. Every time it needs to go to the matrix, take the value and do the calculations. Now it is proposed to change the drop off percentage and reducing the number of calculations to be performed. In this case there is no need to go to the scoring matrix. Another improvement in the algorithm has been carried out in the form of multihit strategy which has resulted in speed/sensitivity tradeoff. Here the user has the option of giving the value of N that determines the number of hit counts required for extension. This step combined with dropoff percentage gives good results and improvement over the existing algorithms.

A major improvement in the algorithm is efficiency as it employs a new approach for curtailing the amount of sequences that proceed for gapped alignment. So this method will work even after the ungapped alignment process is over. This method works better because of the fact that it is not necessary to perform gapped alignment for all the sequences that are coming from ungapped analysis. A middle path approach has been proposed that gives the criteria for reducing the number of sequences going for gapped alignments. Two new algorithms have been developed that have significant improvement over the previously available algorithms and give better results. There is a significant increase in the speed of alignment process without compromising on the sensitivity of results. It deals with a new middle path approach developed for reducing the alignment calculations in BLAST algorithm. This is a new step which has been introduced in BLAST algorithm in between the ungapped and gapped alignments. This step of middle path approach between the ungapped and gapped alignments reduces the number of sequences going for gapped alignment. This results in the improvement in speed for alignment up to 30 percent.

During the course of research work, various points came forth that require the attention of the researchers. The important issues are like integrating the bioscience research with the available computational techniques and methods to deal with the expanding data. There is a need of biomarts to meet up the demand of information required by biologists and computer scientists. Computer scientists have to come up with new algorithms and implementations on the application of soft computing techniques on bioinformatics data. These techniques include parallel processing, distributed processing, genetic programming and neural networks. These have been proposed as future research directions for the researchers working in the area of bioinformatics.