CHAPTER 9

CONCLUSION and FUTURE WORK

9.1. Conclusion

Bioinformatics is an expanding area. Every solution present today extends the horizon and hence new solutions are on the look out for new problems. A lot of new applications is the need of the hour. All these solutions improve the understanding of life by researchers and hence improve quality of life of common man. Some other solutions lead to drug discovery. Some solutions lead to speedy medication.

Bioinformatics algorithms play a crucial role underneath the applications. Research is continuing in the area of algorithms for Bioinformatics. This shows that there is lot of scope for improvement and new algorithms. The fundamental problem in Bioinformatics is LCS identification. LCS identification is part of sequence alignment, homology modeling etc. The mother of all Bioinformatics algorithm being Dynamic Programming methods, parallel algorithms use the technological advancement in hardware. EFPLCS algorithm by the scholar is an Efficient Fast Pruned Parallel Algorithm for identifying LCS which yields more than 70% resource efficiency over the ones like FASTLCS parallel algorithm. This has extended the size of the data i.e. sequence length that could be evaluated with a given set of resources. EFPLCS is scalable for MLCS problems.

More often the Biologists use these ab-initio methods to get a clue about the problem and hence they could decide the direction for further evaluation. The accuracy of the solution is subjective and is acceptable by the Biologists for getting clue. Heuristic algorithms provide faster solution or solution to larger problems at the cost of accuracy. The accuracy depends on the heuristic used which is problem
dependent. To cater to this need of the Biologists' solution, the scholar has developed a **Resource efficient parallel heuristic algorithm** by name SRLCS to identify LCS. SRLCS is heuristic version of EFPLCS. SRLCS algorithm is scalable for MLCS. SRLCS algorithm is comparable to other popular tools like CLUSAL-W, MUSCLE and SSEARCH35.

LCS is the basic step for solution Sequence Alignment. A solution much simpler than heuristic is useful when Biologists or Medical Biologists are working with a specimen sequence and wanting to identify the homology extent with similar ones from homolog database. The simple **SRLCS Model** by the scholar allows one to find the probable length of LCS between the specimen and database sequences. This probable length of LCS is later used in finding the membership of the specimen with homolog database sequences. SRLCS accounts for all the parameters that determine the LCS i.e. the similarity, identity, length of the sequences. The model is trained with training dataset and obtained by regression. It operates on pair wise basis. SRLCS model is validated with t-test using two other test datasets.

The simplistic SRLCS model could be used in places where a case by case comparison is required. The **SRLCS model combined with membership calculator** provides the **membership identification** and is useful to Biologists in many ways. The membership identification is useful wherever a known specimen is under investigation. To the larger society, the scholar has demonstrated the use of Membership Identification in five different scenarios. First scenario is for identification of similar cases of H1N1 in other parts of the country. This was done on protein segments basis, so that a quick medication could be planned. Three other scenarios were used for the membership identification of a specimen protein enzyme in other organisms. Three enzymes Myoglobin, Lysozyme and QNR were dealt with for this purpose. Myoglobin of Vertebrates has better membership with Homosapien Myoglobin than that of invertebrates. Lysosymes themselves have more varieties and
hence difficult to locate near members. QNR is rarely found in near membership in other organisms. The membership value indicates the measure of close homology with the specimen protein. In all cases membership identification was done using “SRLCS Model combined with Membership Calculator”. MSA is required, when the specimen is to be identified as to what it is, from a database. With the known specimen, for any specific query like: “is there an identical case or near case of the instance?” this method is useful.

Curing of database is required while sequencing the samples and creating the database of specimens. The scholar suggests the above membership identification method for curing such databases. This has been applied to a scenario of H1N1 samples and found to be useful.

9.2. Future Work

The future work could be: to apply the SRLCS Model to many problems like DNA Forensic Identification in identifying missing persons, criminals, rare tribes etc on human. Certain problems like a specific enzyme or protein being present in other organisms as identified by the model need to be evaluated for drug design purpose. Also the work could be extended to interesting problems in plants. On the Algorithm area, the EFPLCS and SRLCS algorithms could be applied for finding MLCS. The implementation of MLCS was restricted by the resource availability in the Desktop system used for development work. Thus evaluation of MLCS version of EFPLCS and SRLCS algorithms is a worthwhile future work.

This SRLCS algorithm could be implemented on GPUs or using open MPP for parallel version and compared against DP implementation on GPU. The efficiency is expected to be better.