Abstract

Next Generation Sequencing (NGS) Technologies are relatively high throughput as they produce millions of short reads compared to First Generation Sequencing Technologies that produce relatively fewer but longer reads at orders of magnitude higher cost. Development of new algorithms that can cohere millions of NGS reads while coping up with the sheer volume of the data is presently required to fully harness the potential of NGS technologies for societal benefit, for example, by analyzing sequencing data for predicting, diagnosis and treatment of diseases through personalized medicines and genetic modifications of plants to produce disease-resistant crops.

One of the main applications of NGS in microbiology is the identification of microorganisms such as bacteria, viruses, etc., in a sample. However, a large number of microorganisms have not yet been discovered and hence, their reference genome is not available. Therefore, assembling the entire genome *de novo* remains the most difficult problem for biologists and bioinformaticians. While NGS reads present new challenges because of their quantitative and qualitative nature, they provide new opportunities for assembling and identifying new organisms. Thus, the high throughput provided by NGS technologies can be argued to be an advantage from a computational perspective. In this thesis, it is postulated that the complete overlap information of a NGS read is not required for performing assembly. It is expected that comparable or improved results may be obtained by processing a limited amount of overlap information. Moreover, working with small overlaps avoids time-consuming string operations. This becomes especially relevant when dealing with mammalian size genomes size as human or mouse and large repeat rich model plant genomes like Arabidopsis thaliana. This thesis presents a novel model for assembling NGS reads and an assembler based on it.

In Chapter 2, a novel assembly model based on the above hypothesis is developed and an assembler that works in this model is also developed. Additionally, assembly of ideal reads which are error-free, forward oriented, uniformly distributed over the genome was performed. The results indicated that the hypothesis was valid for ideal data.

In Chapter 3, the problem of error correction in NGS reads is addressed on real datasets. An algorithm tuned to the requirements of the assembler (developed in Chapter 2) was developed. The results indicate improved or comparable performance to contemporary algorithms on real datasets obtained from NGS.

In Chapter 4, the error correction algorithm developed in Chapter 3 is used to correct errors in real NGS reads. The assembly algorithm developed in Chapter 2 is extended to handle corrected NGS reads of 50 bp (base pair) length. This is followed by extending the assembly algorithm to work with reads of 75 and 100 bp lengths. We also showed

that our assembly algorithm is able to handle NGS dataset for large repeat rich model plant genome Arabidopsis thaliana. This indicate that the ideas proposed in the thesis, as they are, can handle genome of such a scale.

Finally, in Chapter 5, the problem of unknown orientations of NGS reads is addressed. For this, a force-based labelling algorithm was developed that assigned orientations to the assemblies and their constituent reads using the assembly overlap graph of the assemblies reported in Chapter 4. Lastly, the assembler was used to generate fewer and longer assemblies using these labelled assemblies. The results indicate that the hypothesis was valid for NGS data.

In conclusion, this thesis reports a novel computational method to analyze datasets with high redundancy. The ideas presented here may also apply not only in genomics but also in problems dealing with a large amount of redundant data.