DEPARTMENT OF BIOTECHNOLOGY
COLLEGE OF AGRICULTURE
JUNAGADH AGRICULTURAL UNIVERSITY
JUNAGADH

Name of Student	**Major Guide**
Ms. Shefalee Kishorebhai Parmar	Dr. M. K. Mandavia

DE NOVO SEQUENCING OF ANCIENT SEED SPICE CELERY (*Apium graveolens* L.) AND DEVELOPMENT OF MICROSATELLITES

ABSTRACT

Key words: Genome sequence, celery, *de novo*, assembly, contig, KEGG, SSRs.

Celery (*Apium graveolens* L.) is one of the most ancient and important seed spice crop. Celery belongs to the Apiaceae family. It is an excellent source of antioxidants and beneficial enzymes, in addition to vitamins and minerals such as vitamin K, vitamin C, potassium, folate and vitamin B6. Considering the importance of this spice plant, various biochemical studies has been carried out and which are still going on including traditional and modern approaches. There is limited genomic data information for the improvement of celery. So, there was a deficit genome data base for the breeder to get the genetic information of celery for the variety improvement and breeding program. Therefore, the present research was undertaken to sequence the genome, to analyze the raw data using bioinformatic tools (CLC workbench v9.5.4) and to develop a set of large number of SSRs from genomic libraries *in silico* and thereafter to validate SSRs. For this study 3 genotypes of celery were used namely, Ajmer Celery-1, Amy Vishnagar and Tall Utah.

In the present study, Ion Torrent Genome Sequencing (Ion S5) technology was used to generate celery draft genome using genotype Ajmer Celery-1. For the measurement of the genome size, flow cytometer (Accuri C6) was used, and genome size of the celery was found 3.94 Gb approximately.

Two runs were carried out in this research study. The data produced after the first run was 6.01 Gb. The second run was preceded by using barcodes for specific samples and that generated 3.3 Gb data. Raw data (Reads) from sequencing (NGS-PGM) was assessed through CLC work bench (Version 9.5.4), in which per-sequence analysis carried out. Initially, the report which was generated after the sequencing of the celery genome comprised of 27,361,042 bp with 197.8 bp as the average read length. Later, trimming process generated a trim report describing the number of reads to be 27,218,172 bp with average read length of 194.4 bp. The *de novo* assembly yielded assembled reads of 441,399,737 and number of contings was 827,971. In the assembly the N25, N50, N75 contig size was 1014, 591 and 395 bp respectively. In *de novo* assembly, 827,971 number of contigs were generated. Somehow, it is a tedious job to perform annotation 827,971 contigs. Therefore, annotation for only selected

contigs was carried out. Contigs with size more than 5000 bp were filtered out and used for the annotation (375 sequences). Blast2GO tool was utilized for the functional annotation and validation of these sequences.

Total 375 sequences were functionally annotated out of which 374 showed positive InterProScan and 355 got Blast hits. 351 and 290 sequences were mapped and annotated respectively. Sequence similarity was also tested based on the protein domain conserved region through InterProScan. 320 sequences showed positive InterPro result while 50 sequences did not show any InterPro results and 130 sequences was scanned with GOs. Mapping database distribution of celery genome showed the highest similarity with UniprotKB (375,000 sequences) followed by Saccharomyces Genome Database (SGD) (85,000 sequences) and The Arabidopsis Information Resources (TAIR) (75,000 sequences). Total 5032 GO IDs were found which were grouped into biological process, cellular components and molecular function. The assembled sequences of celery were divided into six main classes of enzymes. Among all the sequences, 22 sequences were grouped into Oxidoreductases class followed by Transferases class (66 sequences), Hydrolases class (95 sequences), Lyases class (17 sequences) Isomerases class (11 sequences) and Ligases class (07 sequences). In present study KEGG analysis gave 76 pathways.

The SSR primers were designed using BatchPrimer3 (version 1.0) online software. For the identification of SSRs, FASTA file of 375 contigs were used. Total of 100 primers were used, 25 each from celery, ajwain, fennel and dill. 68 SSRs primers were amplified and gave a total of 101 bands in 3 varieties of celery (Ajmer Celery-1, Amy Vishnagar and Tall Utah). The percent polymorphism obtained for SSR primers were ranged from 0% to 100% with an average value of 43.13% per primer. The polymorphic information content (PIC) was calculated for each primer and it was ranged from 0.37 to 0.66, with an average value of 0.19 for each primer. The SSR primer index (SPI) differed from 0.74 to 1.98 with an average value of 0.45.