Whole Genome Assemblies of the Drosophila and Human Genomes.

GENE MYERS
Celera Genomics

Shotgun sequence assembly is a classic inverse problem: given a set of segments randomly sampled from a target sequence, the problem is to reconstruct the target. Early programs for this problem assisted a user by finding potential overlapping segments which were then assembled by hand. As the programs became progressively more sophisticated the problem was completely solved by the software but still followed by a manual curatorial pass by the users. Until 1995 it was believed that the practical limit on the size of problems that could be solved was on the order of 30 to 50Kbp, due to the intrinsic difficulties posed by repetitive sequence in the target. In 1995 the assembly of a whole genome shotgun dataset for H. Influenza dispelled the notion of such a barrier. While the process involved significant human curation and bacterial genomes are less repetitive than those of higher organisms, it still portended an economy of effort unmatched by the more laborious map-based approaches then being pursued for large genomes. In 1996, Weber and Myers proposed a whole genome shotgun approach for the human genome suggesting a protocol that involved sampling several individuals in order to simultaneously obtain polymorphism information. Critics claimed that the computation would involve an impossible amount of computer time, that the size and repetitiveness of the genome would confound all attempts at assembly should sufficient computer efficiency be achieved, and that even if an assembly was produced it would be of an extremely poor quality and partial nature.

In 1999 the informatics research team at Celera produced an assembly of the Drosophila genome from a whole genome shotgun data set consisting of 3.2 million reads, 72% of which were paired-end reads from 2Kbp and 10Kbp inserts in a 1 to 1.32 mix. The assembly consisted of completely ordered and oriented contigs covering an estimated 97.2% of the genome with only 1630 gaps of average size 1,415bp. The smaller gaps were PCR closed by the Berkeley Drosophila project in a three month period following the publication of the assembly, and the remaining gaps closed in the ensuing 6 months. The assembly is consistent with STS maps and physical clone maps and was compared against 24% of the genome independently sequenced by other groups. The sequence level comparison revealed that the sequence is better than 99.998% accurate within non-repetitive regions of the assembly and 99.62% accurate within repetitive constructs. The basic conclusion is that whole genome assembly is not only feasible but produces a high-quality result that requires comparatively little finishing work.

In this talk, we will cover the approaches to sequencing whole genomes, illustrate the key computational steps of Celera's whole genome assembler in an attempt to explain what the critics didn't understand, and describe our current strategies and progress towards a penultimate assembly of the human genome.