COMPUTATIONAL TOOLS FOR NEXT-GENERATION SEQUENCING APPLICATIONS

FRANCISCO M. DE LA VEGA

Applied Biosystems, 850 Lincoln Centre Dr. Foster City, CA 94404, USA

GABOR T. MARTH

Department of Biology, Boston College, 140 Commonwealth Avenue Chestnut Hill, MA 02467, USA

GRANGER SUTTON

J. Craig Venter Institute 9704 Medical Center Drive, Rockville, MD 20850, USA

Next generation, rapid, low-cost genome sequencing promises to address a broad range of genetic analysis applications including: comparative genomics, high-throughput polymorphism detection, analysis of small RNAs, identifying mutant genes in disease pathways, transcriptome profiling, methylation profiling, and chromatin remodeling. One of the ambitious goals for these technologies is to produce a complete human genome in a reasonable time frame for US\$100,000, and eventually US\$1,000. In order to do this, throughput must be increased dramatically. This is achieved by carrying out many parallel reactions. Despite the fact the read-length is short (down to 20-35 bp), the overall throughput is enormous, each run producing up to several hundreds of million reads and billions of base-pairs of sequence data. As the promise of these next generation sequencing (NGS) technologies becomes reality, computational methods for analyzing and managing the massive numbers of short reads produced by these platforms, are urgently needed.

The session of the Pacific Symposium on Biocomputing 2008 "Computational tools for next-generation sequencing applications" aimed to provide the first dedicated forum to discuss the particular challenges that short reads present and the tools and algorithms required for utilizing the staggering volumes of short-read data produced by the new NGS platforms. The session also aimed to establish a discussion between the academic bioinformatics community and their industry counterparts, which are engaged in the development of such platforms, through a discussion panel after the oral presentations of original contributed work. Four contributions were selected from the submissions received and accepted after peer review for inclusion in this proceedings volume and are briefly described next

Given the massive volume of data being produced by NGS platforms, data management becomes a major undertaking for those adopting this technology. New file formats with binary data representation and indexed content will be needed as text files are becoming inefficient both for routine storage and data access. The paper of Phoophakdee and Zaki presents a novel disk-based sequence indexing approach that addresses some of the problems of handling large amounts of data. Trellis+ is an indexing algorithm based on suffix-arrays that allows manipulation of sequence collections using limited amounts of main memory, facilitating NGS sequence analysis with commodity compute servers, rather than requiring specialized hardware. This algorithm can enable rapid sequence assembly and potentially other next generation sequence analysis applications.

Another challenge of analyzing NGS output is the alignment of hundreds of millions of reads coming from a single instrument run to a reference sequence in a reasonable amount of time. Traditional heuristic approaches to sequence alignment do not scale well with short-mers and dynamic programming alignment algorithms such as Smith-Waterman requires significant amount of compute time in commodity hardware, needing embarrassingly parallel approaches, or specialized accelerator chips. The contribution of Coarfa and Milosavljevic is a scalable sequence-matching algorithm based on the positional hashing method. Their current implementation, Pash 2.0, overcomes some of the limitations of positional hashing algorithms in terms of sensitivity to indels, by performing cross-diagonal collation of k-mer matches.

Beyond the (re)-sequencing of regions or whole genomes from pure DNA samples, the sheer volume of data that NGS produce should allow, in principle, tackling the more difficult task of sequencing complex or pooled samples. Sequencing of complex samples is of interest in the case of metagenomics, cancer samples, or mixtures of quickly evolving viral genomes, as well as in genetic epidemiology as a way to address the resequencing of the large number of samples that are needed. The paper of Jojic et al. addresses a significant problem in searching for sequence diversity in HIV genomes from patient samples. Since the virus is evolving rapidly in the host and combination therapy could become ineffective if certain combinations of new acquired mutations evolve, the ability to sequence and distinguish between the viral populations could have major therapeutic implications. The authors describe a method that allows recovering full viral gene sequences (haplotypes) and their frequency in the mixture down to a sensitivity of 0.01%.

Finally, the contribution of Olson et al deals with a new application that NGS enables due to the ability to generate millions of reads from a wide range of positions on the genome. In this case the authors present the tools they have developed to identify a class of small non-coding RNAs of recent relevance, the piwi-associated smallRNAs (piRNAs).

The contributions in this volume certainly address some of the "painpoints" of the utilization of NGS in diverse areas of genome research, but further work is needed. We foresee that initial infrastructural developments would be needed to address the basic analytical and data management tasks that were routine for much lower volumes of Sanger sequencing data. This should be no surprise, since a single NGS instrument can generate an amount of sequence equivalent to that of the entire GeneBank in a short period of time. As time passes and those early problems are overcome, we expect more work on application specific analysis tools to address, e.g. genome-wide gene expression, promoter, methylation and genomic rearrangement profiling. We look forward to such future developments.