A Benchmark of Structural Variant Analysis Tools for Next Generation Sequencing Data

Chatzinikolaou Panagiotis, Bioinformatics & Medical Informatics Team, Biomedical Research Foundation, Academy of Athens, Athens, Greece

Makris Christos, Department of Computer Engineering and Informatics, University of Patras, Rio, Greece

Dimitrios Vlachakis, Bioinformatics & Medical Informatics Team, Biomedical Research Foundation, Academy of Athens, Athens, Greece

Sophia Kossida, Bioinformatics & Medical Informatics Team, Biomedical Research Foundation, Academy of Athens, Athens, Greece

ABSTRACT

In language of genetics and biochemistry, sequencing is the determination of an unbranched biopolymer’s primary structure. A sequence is a symbolic linear depiction, result of sequencing. This sequence is a succinct summary of the most of the sequenced molecule’s atomic-level structure. (Most known is DNA-sequencing, RNA-sequencing, Protein-sequencing and Next-Generation-sequencing)

INTRODUCTION

The process that determines the precise order of nucleotides within a DNA molecule is called DNA-sequencing. A stand of DNA is determined by the order of the four bases (Adenine, Guanine, Cytosine, Thymine) and for this, it uses any method or technology is included. Biological and medical research and discovery has been greatly accelerated cause of the advent of rapid DNA sequencing methods. For basic biological research and in numerous applied fields such as diagnostic, forensic biology, biotechnology and biological systematics, is indispensable to know DNA sequences. Modern DNA sequencing technology achieves the rapid speed of sequencing, while it has been instrumental in the sequencing of complete DNA sequences or genomes of numerous types and species of life. In the early 1970s academic researchers obtained the first DNA sequences using methods based on two-dimensional chromatography, but nowadays, DNA sequencing has become easier and orders of magnitude faster cause of the

DOI: 10.4018/ijsbbt.2013100106
development of fluorescence-based sequencing methods with automated analysis.

The technique that determines an amino acid’s sequence in a protein, the type of the conformation a protein adapts and the extend of complication with any non-peptide molecules, is called protein sequencing. An important tool for understanding cellular processes and easier invention of drags that target specific metabolic pathways, is the discovery of the structures and functions of protein in living organisms. Mass spectrometry and the Edman degradation reaction are the two major direct methods for protein sequencing. Also, we can generate an amino acid sequence from DNA or mRNA sequence by encoding the protein.

The development of the high-throughput sequencing (or else next-generation-sequencing) technologies that parallelize the sequencing process, producing thousands or millions of sequences concurrently, has been driven by the high demand for the low-cost sequencing. Readouts from a variety of DNA preparation protocols to a genome-wide scale, have been fine-tuned from their resolution to single base precision, cause of the new next-generation-sequencing methods. These methods have also transitioned into the sequencing of RNA with result, to include full-length cDNA analyses, serial analysis of gene expression (SAGE)-based methods and noncoding RNA discovery. Novel applications such as the sequencing of ancient RNA samples have been enabled by the next-generation-sequencing, so we notice that the scope of metagenomic analysis of environmentally derived samples have been substantially widened. (Kalb & Moxley, 1992; ten Bosch & Grody, 2008; Tucker et al., 2009) Concluding, NGS has the potential to bring astounding changes in genetic and biological research and to enhance our fundamental biological knowledge.

A snapshot of RNA presence and quantity (Chu & Corey, 2012) from a genome at a given moment I time, can be revealed by RNA-seq (it is also called Whole Transcriptome Shotgun Sequencing) (Fleischmann et al., 1995; Morin et al., 2008), a technology that uses the capabilities of next-generation-sequencing. With NGS, is allowed the base coverage’s increase of a DNA sequence, as well as higher sample throughput, with a result to facilitate sequencing of the RNA-seq can also be used to determine exon/intron boundaries.

After the flourishing of NGS-based methods for genome analysis, the last five years, we are leaded to the discovery of a number of new mutations and fusion transcripts in cancer (Tuch et al., 2010). Researchers could be helped by RNA-seq data interpreting the “personalized transcriptome”, with a result the understand of transcriptome changes happening therefor, or better the determination of gene drivers for diseases like Alzheimer and diabetes. Compared with microarrays, ngs-technology has more advantages to researchers identifying novel and low-frequency RNAs associated with disease process, detecting aberrant mRNA (Twine et al., 2011) and small noncoding RNA expression in disease process (Morán et al., 2012) and the biggest advantage is the low cost with higher throughput (Han et al., 2011).

STRUCTURAL VARIATION

The variation in the structure of an organism’s chromosome is called structural variation (or else genomic structural variation). Microscopic and submicroscopic types, such as deletions, duplications, copy-number variants, insertions and translocations are usually included in SVs and it consists of many kinds of variation on the genome of one species. Typically a structure variation affects a sequence length between SNPs and chromosome abnormality (Feuk et al., 2006). SV is not wilful about frequency of phenotypical effects but it can cause genetic diseases. Researches show that SVs are more difficult to detect than SNPs. There are many kinds of SVs, such as microscopic structural variation (Reich et al., 2006; Gripenberg, 1964; Wyandt & Tonk, 2004), copy-number variation (Sebat et al., 2004; Iafrate et al., 2004; Lupski, 2010; Lam et al., 2010), inversion (Lakich et al., 1993; Bondeson et al., 1995; Tuzun et al., 2005;
Differential Association Rules: Understanding Annotations in Protein Interaction Networks
Christopher Besemann, Anne Denton, Ajay Yekirala, Ron Hutchison and Marc Anderson (2006). Advanced Data Mining Technologies in Bioinformatics (pp. 269-282).
www.igi-global.com/chapter/differential-association-rules/4256?camid=4v1a