Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 9 **Editorial**

What to do about data?

Martin Gollery*

Tahoe Informatics Reno, Nevada, United States; Martin gallery - Email: marty.gollery@gmail.com; *Corresponding author

Received on August 04, 2010; Accepted January 28, 2011; Published February 07, 2011

The cost of sequencing whole mammalian genomes continues to plummet. New sequencing technologies are in development in many locations around the world, promising miraculously low prices and unbelievably short analysis times. When discussing the '$1,000 genome', the question is no longer whether it will be achievable, but when. No matter how cheap sequencing the human genome can be, there is always someone promising better results for less money in a shorter time, with delivery starting in 12-18 months. Now one group even promises that they will be able to deliver a human genome for $30!

Whether the hyperbole surrounding these companies will ever turn out to be valid or not, there is no doubt that we will see an influx of tens or hundreds of thousands of human genomes over the next few years. It is imperative for those of us in the business of dealing with bioinformation that we implement the best practices and methods to deal with this kind of data sooner rather than later. Experience has shown that bad practices, once established, are difficult to shed.

Treating human genomes as if they were simply big sequence files is naïve. Some researchers want to be able to treat 100,000 genome sequences as they used to treat 100,000 EST sequences. This is a recipe for disastrous wasting of resources. Others point out that one could simply save the reference genome and then have all future genomes simply refer to the reference and list variations from that reference. This is vastly superior to a simplistic listing of each genome in its entirety, but is still insufficient in the long term.

To begin with, we need to understand that the situation with sequence data has been a mess for quite some time. The working bioinformaticist has had to contend with a multitude of database and alignment formats, both in text and binary, for a number of years now. No truly flexible format has been available, so nearly every new tool has also had to introduce a new data structure.

The BioHDF project will take a giant step towards solving the format compatibility problem. Currently in development, the BioHDF format and associated tools will provide a flexible solution that can accommodate data that ranges from sequence to alignments to Mass Spec, with the ability to incorporate new data types as they are developed. Compression and an efficient indexing scheme means that the data may be extracted easily from anywhere in the file without having to unzip or search through the entire database. An API and library based on C is being developed to provide functions which allow the import and export of data to and from existing bioinformatics file formats. Wrappers are being generated for those who prefer Perl.

Still, I wonder if there is a better way to capture all of the information generated by the various human genome sequencing efforts. While capturing the variants from a reference genome is great, we might want to assign probabilities to indicate how rare that variant is. Rather than simply indicating that it is a variant, we could see whether it is common or uncommon. Then when a new sample comes in with that variant, we would have some reference to assess the significance of this variant in different contexts. The data would then look less like a FASTA file and more like a hidden Markov model.

Collections of these probabilistic models could help to further enhance the value of large scale genome projects. After all, a position may be considered a 'rare variant' when compared to NCBI Build 36, but may be perfectly common among the population group of interest. The genome of an Asian woman with schizophrenia, for example, may be best compared to a database of similar genomes rather than a generic 'one-size-fits-all' type of reference.

Ultimately, the accumulation of hundreds of thousands of human genomes will be beneficial to those who seek to understand the workings of the human organism, and therefore, to all of us. The only question is, how can we best plan for the abundance of data that will be available in the future and make the most efficient use of it?

Edited by P Kangueane

Citation: Gollery. Bioinformation 5(9): 367 (2011)