Computational solutions for omics data (Summary)

In the area of high-throughput sequencing, data generation and computing power are being increasingly diverged over time such that omics data analyses pose certain difficulties.

One such difficulty is the algorithmic efficiency needed to process large datasets. Genome assembly deals with the generation of a reference genome to which sequences can be analysed against and thus needs to be accurate, fast and have efficient storage methods. Most efficient assemblers are built on the de Bruijn graph as it reduces fragment assembly to the Eulerian path while others use FM-indexing. Both graph-theoretical methods still encounter problems to accurately assemble a large genome such as the human one. Read mapping in next-generation sequencing deals with the mapping of reads to a reference genome and thus needs to incur low running times and computational costs. Most efficient short-read aligners use FM-indexing to pre-process the genome compactly. Hardware accelerated algorithms use parallel dynamic programming, multicore CPUs or cache-oblivious algorithms to increase the software's speed. Large-scale genome sequencing deals with reducing the size of sequencing data for storage and analysis (sequence search). Compression algorithms include reference-based methods for re-sequencing and non-reference-based methods for repetitive DNA segments. Compressive genomics, used since search algorithms are becoming too slow, compresses data such that analysis can be implemented without decompression needing to be done by making use of genomic redundancy.

Another difficulty is data mining for transcriptomics – transcriptome quantification by RNA sequencing (RNA-seq). Identifying cell-specific expression signals within tissue profiles is done using linear algebraic methods which need measured cell type proportions to weight a linear mixed model. When proportions are not available, matrix factorisation or differential geometry methods are used to estimate. Identifying regulatory and phenotypic genes and modules (gene expression analysis) is done using: statistical methods, probabilistic graphical models, and sparse learning. Identifying gene expression alterations in disease (ex. comparisons between tumour cells and normal cells) is done using software which either extends the graphical model or uses a Bayesian network to construct pathways. Methods are not standardized so large-scale application is lacking.

The other difficulty is integrative interactomics whose analyses involve modularity – interactomes/networks being represented as graphs. Analysis of heterogeneous genomic data sets can be done using sub-networks and local clustering to uncover specific modules of interest. Network flow can be used to propagate biological processes or to identify proteins. Cellular networks can also be used for module and pathway analysis though random walk-based approaches; for example the Isorank algorithm. Interactome analysis of disease data sets are due to mutations and variations identified by sequencing certain individuals. They show that the genes may differ but the pathways are generally shared amongst the afflicted. Genome-wide association studies (GWASs) identify these leading to disease gene prioritisation and uncovering of related pathways. The modularity of such genes can be tested using permutation-based approaches. Also, the problem of set cover has been found to be beneficial when considering this heterogeneity.

Genomes, transcriptomes, proteomes, interactomes and methylomes are laboratory generated. Omics analyses' algorithms lead to the correct usage of these in biological areas. All the above areas should be researched extensively as high-throughput technologies are continuously advancing.