OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of
Data Mining in
Genomics & Proteomics

Commentary

Integrative Genomics from Gene Data Mining to Disease Genome Sequence Analyses

Madeddu Urrutia*

Department of Biology and Biochemistry, University of Sheffield, Sheffield, UK

DESCRIPTION

A data mining approach is required at all levels of genomics and proteomics analysis. These studies provide a wealth of information and can quickly generate large amounts of data from the analysis of biological samples from healthy or diseased tissues. Data mining was the easy extraction of potentially useful information from data that was implicit and previously unknown. In recent years, genomic data has exploded. This is due to advances in various high-throughput biotechnologies such as RNA gene expression microarrays. These large genomic datasets are informative and often contain far more information than the researchers who produced the expected data. Such vast amounts of data allow for novel analysis, but make it difficult to answer research questions in the traditional way. Analyzing this vast amount of genomic data poses some unprecedented challenges.

Analysis of high-throughput genomic data requires processing an astronomical number of potential targets, most of which are false positives. For example, a traditional statistical test with a significance level of 5% yields an average of 500 false-positive genes from a 10K microarray study comparing two biological conditions with no actual biological difference in gene regulation. Indeed, if there are a small number of 100 genes that are differentially regulated, then such true differentially expressed genes would be the 500 above without prior information to distinguish between the two groups of genes. The 600 targets identified by such statistical tests are unreliable, and further investigation of these candidates yields inadequate results.

Simply tightening these statistical criteria significance level of 1% or less increases the false-negative error rate and makes it impossible to identify many important true biological targets. As the number of signaling pathways or interaction mechanisms in question grows exponentially, this type of pitfall, the so-called multiple comparison problem, presents new biological is biomarker predictions involving multiple interacting targets and genes. It gets even worse when trying to find a model. Therefore, it is important that data mining techniques effectively minimize both false-positive and false-negative error rates in these types of genome-wide studies.

When analyzing current genomic data, the use of machine learning and data mining techniques has become more attractive as the complexity of the project has increased. Approaches from this area were studied as part of the Genetic Analysis and were motivated by two main starting points. First, assuming the underlying structures of genomic data, data mining can identify them and improve downstream association analysis. Second, we need to develop more machine learning computations to be able to efficiently process today's abundant data.

In the process of discussing the results and experiences of machine learning and data mining approaches, six general messages were extracted. These show the current status of these approaches when applied to complex genomic data. While some challenges remain for future research, significant advances have been made in the integration of different data types and the evaluation of evidence. Searching for data on the underlying genetic or phenotypic structure and using this information in subsequent analyzes proves to be very useful and may be even more useful in more complex datasets. Analyzing complex genomic data is a challenging endeavor that can be tackled using machine learning and data mining techniques. What all these methods have in common is that they search for patterns in the data.

To distinguish between machine learning and data mining, data mining is described as the process of extracting useful information from data. In contrast, machine learning can be seen as a set of methodological tools for extraction. Therefore, data mining involves data selection, preprocessing, and transformation until the actual application of machine learning techniques to create the model, and the model is interpreted and evaluated. Therefore, machine learning can be seen as a particular aspect of a larger class of data mining techniques that

Correspondence to: Madeddu Urrutia, Department of Biology and Biochemistry, University of Sheffield, Sheffield, UK, E-mail: Madedduurrutia@yahoo.com

Received: 10-Jan-2022, Manuscript No. JDMGP-22-15246;**Editor assigned:** 12-Jan-2022, PreQC No.JDMGP-22-15246 (PQ);**Reviewed:** 24-Jan-2022, QC No. JDMGP-22-15246;**Revised:** 27-Jan-2022, Manuscript No. JDMGP-22-15246 (R); **Published:**31-Jan-2022, DOI: 10.4172/2153-0602.22.13.246.

Citation: Urrutia M (2022) Integrative Genomics from Gene Data Mining to Disease Genome Sequence Analyses. J Data Mining Genomics Proteomics. 13:246

focuses on algorithms for automatically detecting patterns in data.