Data Science Reveals Unique ALS Gene Associations

Standard Bioinformatics Methods Fail to Identify Gene Expression Cohort

Amyotrophic lateral sclerosis type 4 (ALS4) is a rare form of ALS featuring juvenile onset, lack of bulbar involvement and indolent course (most patients survive for more than 50 years post onset) which distinguish it from classic ALS. Netrias, in collaboration with clinical researchers at the NIH National Institute for Neurological Disorders and Stroke (NINDS), applied advanced Data Science approaches to the analysis of RNA sequencing data from ALS4 and control patients to identify differential expression patterns contributing to the disease.

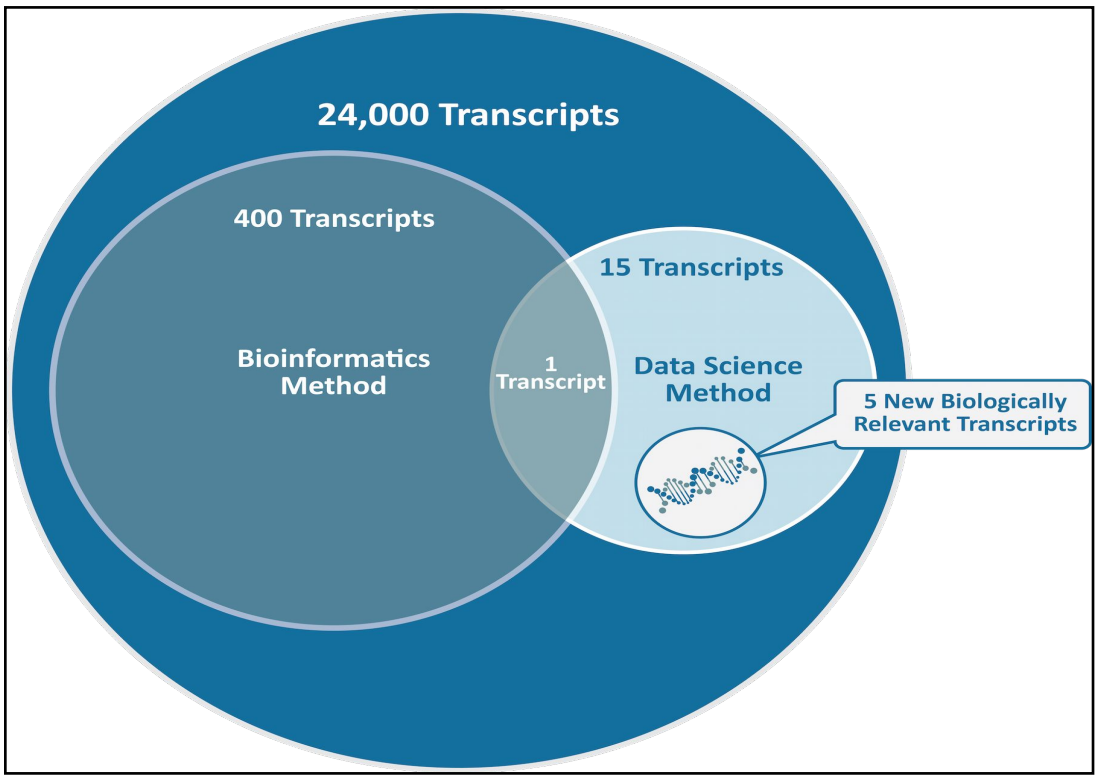

Figure 1: Data Science approach yields distinct and separate results as compared to conventional Bioinformatics approach

Using this approach, Netrias identified 15 candidate gene transcripts that were separate and distinct from the ~400 candidate transcripts identified using conventional Bioinformatics approaches (Figure 1). Of our 15 candidates, five genes were determined by our collaborator to be biologically interesting, and six others merited further investigation to determine their relationship to ALS4.

Netrias, LLC
4800 Hampden Lane, Suite 200
Bethesda, MD 20814
240.482.3700

Conventional Bioinformatics Does Not Fully Enable Discovery

In the conventional Bioinformatics process performed by our collaborator, genes were aggressively filtered to reduce the number to validate experimentally. First, a hard threshold was applied on the Reads Per Kilobase of transcript per Million (RPKM) mapped reads. RPKM is a standard normalization process for post RNA Sequencing [1]. Our collaborator established a hard RPKM threshold by visual inspection using a genome viewer, selecting the minimum threshold of gene expression that showed the best alignment with a reference genome.[1] A second filtering step removed genes which did not occur in at least 75 percent of the samples. The remaining gene expression levels were averaged, fold ratios were calculated, and a final filter was applied that limited the resulting expression levels to two-fold or greater. This typical approach was replicated by Netrias, and identified approximately 400 candidate genes that required further biological validation. With this large number of gene candidates, the manual process of validating against empirical evidence and the scientific literature would be extremely time consuming. Clearly, an alternate data analysis method was required to reduce this result set to a manageable and meaningful number of candidates genes with a high correlation to the ALS4 disease.

Netrias' Data Science Process - Data Census and Insight Generation

Netrias took a data science approach to analyze ALS4 gene expression. Using the same RPKM normalized data, we performed a two step proprietary process: Data Census and Insight Generation. Our results consisted of 16 gene transcripts, 15 of which were distinct from our collaborator's 400, and one that was initially identified among the 400. Five of the 15 gene transcripts were rapidly identified by our collaborator to be biologically interesting, and six others required further investigation to determine their relationship to ALS4.

Netrias enables data-driven decisions with no prior domain context through a two-phased proprietary process. We perform a Data Census to extract emergent properties from the dataset under investigation, and those properties inform our Insight Generation process to derive meaningful results. This allows for an unbiased data reduction process compared to aggressive filtering with ad-hoc thresholds that ignores the full context and underlying properties of the data.

Data Census, an approach that Netrias originally derived to extract properties of Big Data, reviews all of the data and extracts properties such as first order statistics, trends, and data types (categorical or continuous). This informs statistical analysis parameters that generate valuable results from the data [2-4]. In the research conducted on the ALS4 dataset, Data Census revealed a large degree of expression variability. This property led us to conclude that the conventional bioinformatics analysis that averaged the gene expression levels across the cases did not adequately account for the large variation across the three ALS4 patients. We therefore individualized, rather than averaged, the expression levels of every case within our analysis to produce an unbiased fold change ratio in the Insight Generation stage.

[1] RPKM threshold is selected based on testing a number genes that cover a small discrete range of values to identify the threshold cutoff where the sample best aligns to a reference genome. An alignment tool, such as Integrated Genome Viewer (IGV), can align the bam files against the reference genome. The threshold is then selected by visually inspecting the genes that were selected at the different RPKM that best align to the reference genome. Any expression below that value is discarded.

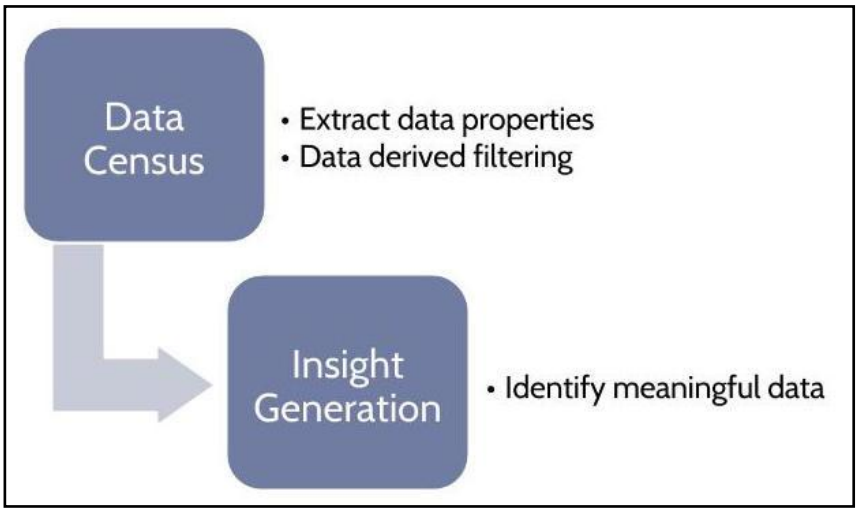

Figure 2: Netrias Data Science Process

Insight Generation leverages unsupervised Machine Learning (ML) techniques to find unknown unknowns [5-8]. It reliably identifies unknown relationships because we build adaptable, data-driven models as compared to rigid, rule-based models. ML has revolutionized a variety of domains including finance, cybersecurity and social media. [5]

In the research conducted on the ALS4 dataset, we represented gene expression levels in a three-dimensional gene space (one dimension per ALS4 patient). Each point in gene space encoded a fold ratio from each patient as determined by comparing patient expressions to the average control. We computed first order statistical parameters during Data Census such as median expression to identify standard deviation below 100 across the disease patients that informed our first-round, data driven filter. We then selected the transcripts that shared at least 20% similarity across the disease patients. The thresholds on the standard deviation and similarity ratio were chosen to limit transcript count, and can be iteratively applied to expand the list as required. A list of prioritized candidate genes based on this expression analysis were then returned to our collaborator.

Next Steps: Extend the Analysis to Include More and New Data Types

In the next stage of our collaboration, we will extend our analysis to include datasets with larger sample sizes, validate the biological impact associated with our initial findings, investigate differential tissue expression, and further correlate our findings with other ‘omics data that help differentiate the ALS4 disease condition. Through this process, we will adapt our analytical approaches to generate the most interesting, biologically-relevant results for our scientific partner.

Data Science Approach Decreases Time to Value, Increases Precision, and Produces Unique Data-Driven Results

The chief benefits of our approach as compared to the more conventional Bioinformatics approach are time to value, precision, and unique results.

Our Data Census method allows us to approach a dataset with no context whatsoever, freeing a collaborator to focus on other tasks while we generate data properties. This focuses our attention on the most salient features in the dataset. Furthermore, we can perform this phase rapidly and independently, with a minimum amount of attention from our collaborator.

The Insight Generation approach finds unknown unknowns, the relationships or findings that are hidden and unique to the dataset under analysis. It identifies unknown relationships by building precise, adaptable, data-driven models as compared to rigid, rule-based models. These data models provide the ability to examine multiple parameters leading to unique data insights. Insight Generation can scale across technology stacks and platforms - from local machines to High Performance Computing (HPC) and Cloud - leveraging any single or ensemble of ML algorithms.

Citations

1. Li, P., Piao, Y., Shon, H. S., & Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of Illumina high- throughput RNA-Seq data. BMC Bioinformatics, 1–9. http://doi.org/10.1186/s12859-015-0778-7
2. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007
3. Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. National Science Review, 1(2), 293–314. http://doi.org/10.1093/nsr/nwt032
4. Wu, Xindong, et al. Data mining with big data. IEEE transactions on knowledge and data engineering 26.1 (2014): 97-107.
5. Bryant, R.E., Katz, R.H., Lazowska, E.D. (2008) Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. Computing Community Consortium. Version 8. http://www.datascienceassn.org/sites/default/files/Big%20Data%20Computing%202008%20Paper.pdf
6. Leung, M. K. K., Delong, A., Alipanahi, B., & Frey, B. J. (2016). Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. Proceedings of the IEEE, 104(1), 176–197. http://doi.org/10.1109/JPROC.2015.2494198
7. Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. Big data analytics in bioinformatics: A machine learning perspective. arXiv preprint arXiv:1506.05101, 2015.
8. Carter, H., Hofree, M., & Ideker, T. (2013). ScienceDirect Genotype to phenotype via network analysis. Current Opinion in Genetics & Development, 23(6), 611–621. http://doi.org/10.1016/j.gde.2013.10.003
9. Rost, B., Radivojac, P., & Bromberg, Y. (2016). Protein function in precision medicine: deep understanding with machine learning. FEBS Letters, 1–26. http://doi.org/10.1002/1873-3468.12307