Beyond Mendel, beyond Koch: Unravelling multifactorial causation in complex diseases with AI and (Meta-) genomic big data.

Andreas Henschel*

Khalifa University of Science and Technology, Abu Dhabi, UAE

Accepted on February 01, 2018

Introduction

Many diseases are not monocausal. This holds true for both genetic and infectious diseases. The advent of whole genome sequencing, fueled by drastically dropping DNA sequencing costs, generated huge expections to revolutionize etiology, diagnosis, treatment and drug discovery.

Indeed, many efforts have been made to unravel the causative genetic variants in Mendelian diseases, i.e., those disorders attributable to a single gene. They were the lowest hanging fruits in the search for heritability. However, the Human Genome project fell short of some initial -unreasonably high- expectations, as most reported variants confer only minimal disease risk. Researchers have thus increasingly looked at the possibility that multiple causative variants could explain a phenotype.

The duality in the search for causation in genetic and infectious diseases is striking: the equivalent to a Mendelian disease is an infectious disease caused by a single pathogen. The predominant paradigms to diagnose infections have been defined by Koch's postulates. They demand that the causative microorganism must be found and isolated in abundance in all affected patients but not in healthy ones. This approach has its limitations, as only 1% of microbes can be cultured. Even if all pathogens could be isolated, the recent surge in microbiome research has brought up a more encompassing picture: dysbiosis (the disbalance of the microbiome) has been reported to happen in a huge variety of ways, not attributable to a single operational taxonomic unit (OTU), whether this may be a strain, species or even genus. Changes in the environment, such as acidity, salinity, diet and antibiotics can incur complex ecological processes. Metagenomics, deploying either shotgun or marker gene (16S rRNA) sequencing, provides a glimpse into the entire spectrum of possible explanations. Identification of (nearly) all potential agents contributing to infection is the equivalent of calling (nearly) all variants in Genome Wide Association Studies (GWAS). The parallels of the two realms thus extend to the recent paradigm shifts in data acquisition: universal instead of pathogen-targeted primers are the Metagenomics equivalent of going from sequencing single genes or gene panels to genome wide genotyping or whole exome/genome sequencing. The more comprehensive the data acquisition, the less one is hampered by preconceived ideas. On the flipside, broadening the view comes with a price: an overwhelming computational complexity. Already most GWAS lack the statistical power to test for multiple factors due to sample size using conventional statistical approaches.

From an Artificial Intelligence perspective, while not directly resolving causation, the data in the two respective realms lend themselves to be formalized as (phenotype) classification problems using Supervised Machine Learning. A sample is formally quantified as a feature vector consisting of genetic variants/OTUs, labeled by phenotype.

The central precondition for this approach is to overcome the “curse of dimensionality", i.e. to keep the dimension of the feature space in an appropriate ratio with the sample size. In both realms, deep sequencing creates initial feature space dimensions in the millions, i.e, several orders of magnitude higher than the number of available samples. With so many features, chances are that one or few of the features explain the phenotype simply by coincidence, a problem refered to as p-hacking and related to overfitting. Dimensionality reduction techniques such as Feature Selection (optimally selecting feature) are well-established, but are intrinsically hard problems (NP-complete, in Computer Science terms). Despite these issues, classic Machine Learning has already been successfully applied in phenotype prediction, through heuristic reduction algorithms (manual and automated) and overfitting solutions (as in Support Vector Machines).

In this context, two remarkable recent developments open further opportunities: the increased data intensity of Life Sciences and the co-evolving next generation of AI algorithms known as Deep Learning. Its striking advantage is its ability to maintain a steadily rising performance curve with growing data volume in contrast to classic Machine Learning techniques, which eventually plateau.

Conveniently, the current decade has marked an era of (Meta-) genomic Big Data sets. Worth mentioning are two UK-based efforts - a recent meta-analysis of 98 Type 2 Diabetes GWAS (which combines 1.25 Million samples) and the “100.000 Genome" project. In the realm of Metagenomics, similar scales are reached by the MG-RAST database, the Human Microbiome Project and the Earth Microbiome Project. Datasets of this size offer fertile ground to apply Deep Learning algorithms such as Deep Neural Networks (DNNs). On the algorithmic side, DNNs excel due to a series of innovations that helped to overcome the Vanishing Gradient problem during training, thus enabling deeper networks. Finally, Deep Learning has profited from improved utilization of hardware such as GPUs.

AI is still dealing with a few challenges in addition to the abovementioned overfitting and feature selection problem. While classification models have reached high accuracies (as demonstrated by cross validation), AI methods still lack explanatory power, a problem that is even exacerbated in Deep

Neural Networks. Their complexity doesn't allow insights on how classification decisions are drawn. One step towards better explanations is gauging the information gain of features during classification. For example, Random Forest Trees include feature ranking which highlight the most informative features. Artificial Neural Networks can be simplified through "dropout", i.e., the omission of "neurons". A related open challenge in Feature Selection is to deal with Linkage Disequilibrium and respectively co-occurrence of OTUs in microbiomes, i.e., discerning causation from correlation.

In summary: Deep Sequencing meets Deep Learning. The arrival of very large scale genomic and metagenomic datasets has placed AI in a promising position to address complex diseases. The full potential of these techniques will be unlocked, if the “curse of dimensionality" can be handled. Phenotype prediction is already a tangible task. Further advances in explanatory AI can be expected, e.g., through model simplification such as the “dropout" technique and more rigorous feature selection. This ultimately will narrow the search space for causes of multifactorial diseases.

***Correspondence to:**

Andreas Henschel
Khalifa University of Science and Technology,
Abu Dhabi,
UAE