Innovative goals

Prediction of human pathologies based on genetic information is a cutting-edge target for personalised medicine. However, there is a number of hurdles hampering the development of genetic diagnostics. These, among others, include:

1) High dimensionality of genetic feature space dominated by non-informative features. Indeed, each of us carry millions of mutations inherited from our parents and every day 37 trillion of our cells acquire trillions of new mutations (Milholland et al., 2017). However, most of these mutations have little to no impact on human health and only few may lead to disease. Identification of novel mutations in every cell for early disease (cancer) diagnostics deems practically impossible at the current state of technology. Nonetheless, a simple yet elegant reasoning led to conclusion that over 50% of cancer cases could be explained by new mutations generated during each cell division, while hereditary mutations explain only a minor proportion of cancer risk (Tomasetti et al., 2017). This argument on itself questions utility of genetic cancer risk diagnostics, except for few cases of monogenic predisposition to cancer when association of a mutation with particular cancer is well established.

2) Inability of genetic models to account for environmental factors affecting actual human state and health. For example, it is well known that smoking, obesity, environmental pollutions, *etc.* increase risk of cancer, but this information has to be collected and factored in separately into genetic models. This, in turn, lowers the power of “pure” genetic diagnostics.

3) Last, but not least, collection of genetic information required for genetic diagnostics could potentially deanonymize individual’s identity.

Alternatively, one might take a look at gene function for diagnostics of human pathologies. In brief, central dogma of molecular biology in its simplified formulation states that genetic information flows from nucleic acid to nucleic acid (from DNA to RNA) and then from RNA to protein, and in organism proteins exert most of biological functions. Thus, to assess normal or pathological gene function researchers commonly estimate copy number of either RNA or protein molecules synthesized from a given gene and refer to this as differential gene expression analysis. The idea that genes respond to environmental or physiological signals by changing their expression levels (RNA/protein copy number) dates back to 1960 seminal paper by François Jacob and Jacques Monod: "L'opéron: groupe de gènes à expression coordonnée par un opérateur" (Jacob et al., 1960). Thus, diagnostic models based on the analysis of gene expression (function) account for environmental conditions and, as a result, may reflect better actual human health state. To that, analysis of gene expression does not uncover sensitive human genetic information, thus, protecting personal rights. However, diagnostics of human pathologies based on gene expression profile is not without its own caveats:

1) Although dimensionality of gene expression feature space is several orders of magnitude lower than that of genetic features, it is still large (there are thousands of active genes each of which potentially might be involved in the development of human diseases).

2) Gene expression is a stochastic process and there is a significant amount of variation in RNA and protein copy numbers between individuals, also known as gene noise (de Jong et al., 2019). In cancer patients, gene noise is anticipated to be even higher than in healthy individuals due to heterogeneity and aneuploidy (abnormal chromosome and gene copy number) of cancer cells. As a result, this complicates molecular diagnostics, identification of cancer etymology and discovery of therapeutic gene targets.

3) Finally, it has to be noted that practically quantification of gene expression by RT-PCR (reverse transcription polymerase-chain reaction) of RNA molecules is most reproducible and cost-effective approach, but it does not allow to count RNA copy-number in absolute terms. Instead, in RT-PCR gene expression is estimated in relative terms by calculating the ratio of a gene of interest (target gene) to a reference gene (a control gene, which is assumed to be unperturbed by pathological or other conditions). Unfortunately, this only increases uncertainties and complicates the development of diagnostic models.

In spite of these considerations, attempts to associate specific mutations in genes or altered gene expression levels with specific diseases are being actively propagated by researchers and hyped in the media. In part, this trend is wormed up by pharmaceutical companies, who seek for new therapeutic targets (genes) to design new sellable drugs for patients. *Per se* we don’t see any problems with this strategy, but we feel that such trends create observational bias. In other words, we limit our search to only “low hanging fruits”.

To circumvent abovementioned concerns, we developed a novel approach to gene expression analysis and training of diagnostic and prognostic human pathology models. First, it has to be noted that most of human pathologies affect expression of not just one but rather of many genes. Second, it is then easy to realize that disease might have an impact on many parts of entire gene network. Thus, instead of searching for the specific genes we might try to quantify changes in stoichiometries (proportions) of gene network constituents. This led us to an idea of using gene expression stoichiometries to develop diagnostic and prognostic models for human pathologies. In principle, our “wholistic” approach results in the following key advantages:

1) Significant reduction in stoichiometry signatures models’ dimensionality.
2) Reduction in noise for stoichiometry signatures. As a result, reduced noise in feature space simplifies models training and improves their cross-validation accuracy.
3) Models trained on big data generated by high-throughput RNA-sequencing or microarrays technologies potentially can be transferred to low cost platforms, such as RT-PCR.
4) Last, but not least, our approach may uncover novel therapeutic targets based on analysis of stoichiometry changes in constituents of gene network.

Below, we illustrate the power of our approach on several examples concerning predictive diagnostics of the efficacy of breast and multiple myeloma anti-cancer chemotherapies.