Deep learning-based cancer patient stratification

Abhinav Mishra, Jule Brenningmeyer, Maike Herkenrath
Department of Mathematics & Informatics, FU Berlin

Introduction

To develop a cancer drug, in addition to direct medical costs in the development, the incurred cost are amounting to 2.7 billion dollars which is embedded within a long, and time-consuming process. This process includes pre-clinical research, clinical trials, approval process, and prescription guidelines. A simplistic, and narrow focused approach in the process of drug development is the most recurring reason for inefficiencies, e.g. In therapeutic decisions, not taking the whole genome into consideration which results in information loss. **ARCAS** project aims to improve the situation by data integration, and analysis that will create an assumption-free approach for modelling clinical variables. The multi-omics integrative deep learning framework **ARCASOnco** models variables for survival, and drug response with interpretation to understand molecular mechanisms or pathways corresponding to their latent factors.

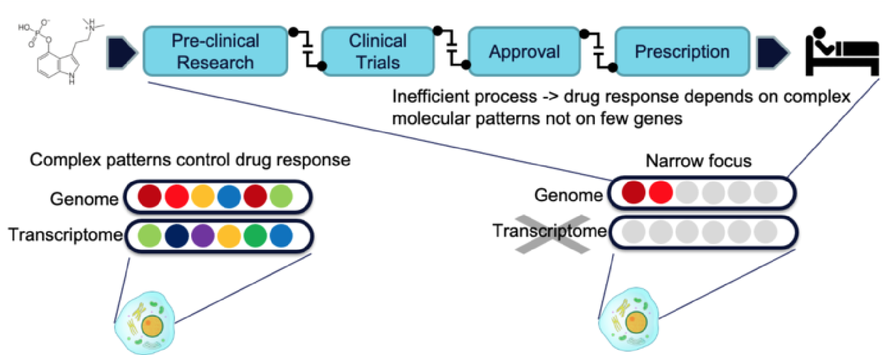

Figure 1: Drug development process. Use of genomics data

Methods

ARCAS have developed a framework which uses deep learning to integrate any kind of omics data and discover molecular patterns, or so-called latent factors. Data integration for cancer omics can deliver molecular patterns predictive of clinical variables.

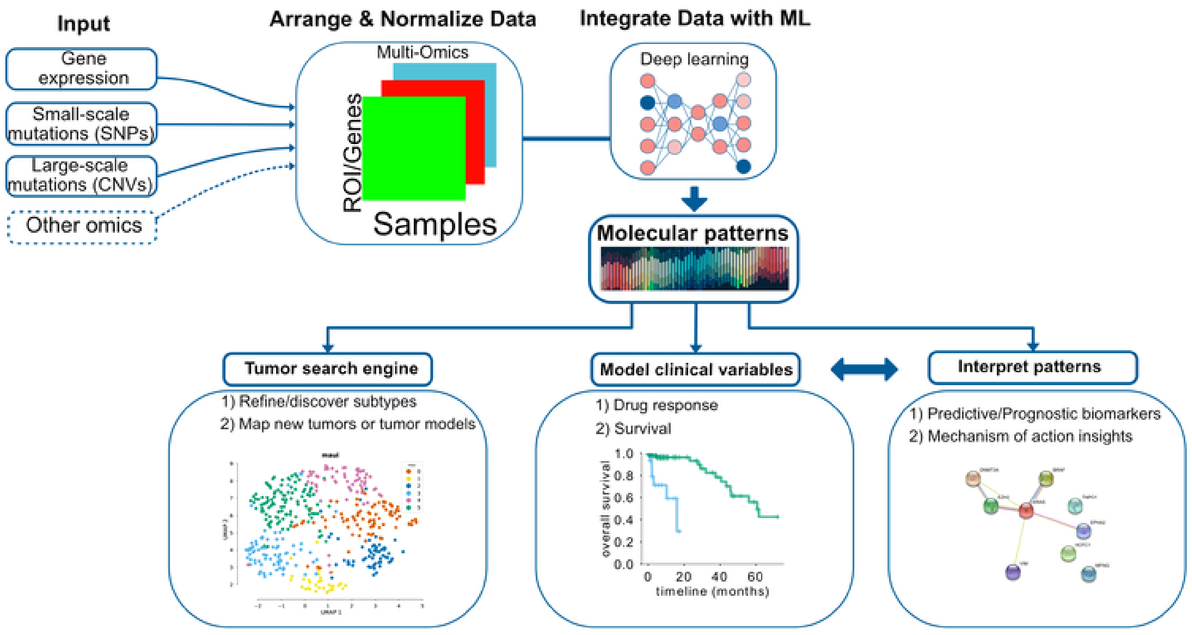

Figure 2: Data integration for cancer omics can deliver molecular patterns predictive of clinical variables.

Results 2

Further refining the subtypes was achieved by applying a clustering algorithm. Six clusters were find. Compared to the plot on right side, CMS2 was separated into two clusters. In terms of survival, this actually makes a lot of sense. Below, we show survival curves of these new two clusters, which are very different. Therefore, it is justified to break-up CMS2 to two subtypes.

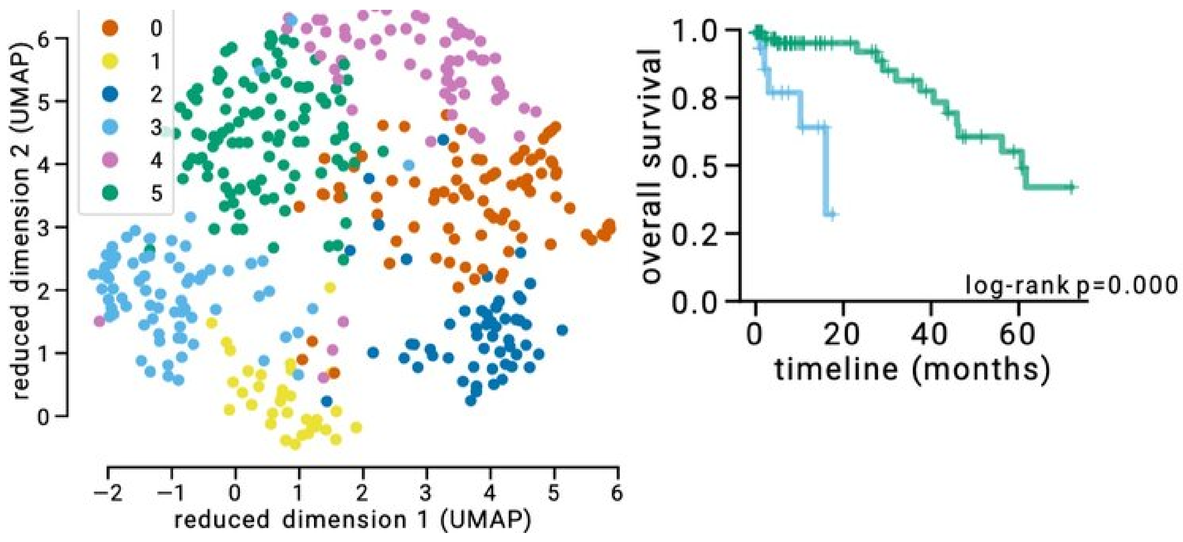

Figure 4: Refined subtypes for colorectal cancer. Green and Blue dots represents CMS2 subtype, however separating this subtype into 2 subtypes as suggested by latent factor clustering makes more sense in terms of different survival characteristics of these two groups.

Results 1

If molecular patterns, i.e. latent factors, contain relevant information, it should be possible to predict the CMS from them. The receiver operating characteristic curve, or ROC curve, shows that latent factors are able to predict CMS subtypes (left). The developed method was compared with other methods and showed in all cases a higher prediction accuracy. The 2D projection of latent factors was color-coded based on the CMS status (right). Each dot is a primary tumor colored by the CMS status. Separation of colors can be seen, which means there is information in latent factors about CMS status.

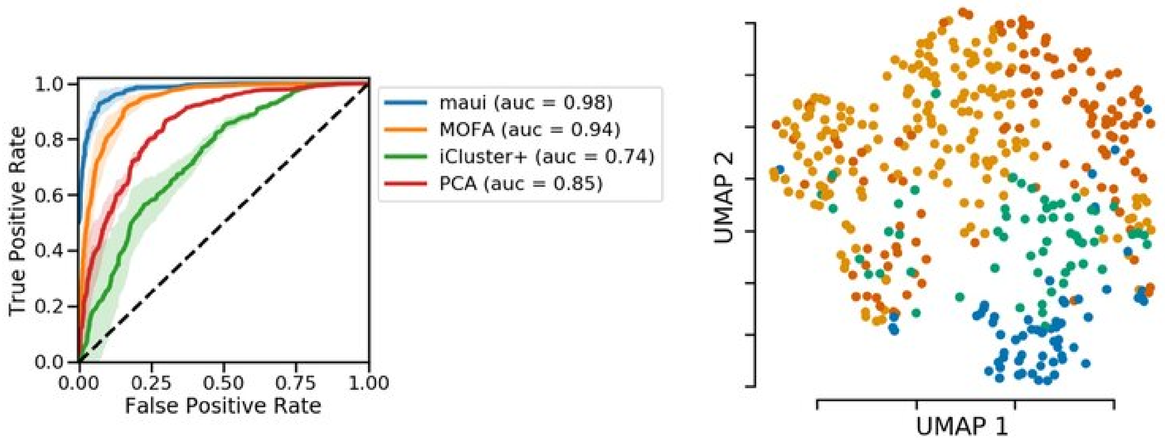

Figure 3: Predicting subtypes using latent factors obtained via deep learning is more accurate. Left, accuracy in comparison to other toos. Right, Representation of colorectal tumors by reducing latent factors to 2D.

Results 3

The method works in any data set that has multi-omics information, including tumor models, such as cell lines, PDX or organoids. In fact by using the method cell lines, PDX, and primary tumors can be integrated.

The method was used on a cancer genome atlas data sets that had at least 100 samples. In figure 5 the improvement of C-index, which is a measure of survival prediction accuracy, is shown. Furthermore the figure shows what happens when predicting survival just by using clinical variables, such as age, gender, and tumor stage - in comparison to clinical variables + latent factors. In many cancers the accuracy metric is pushed to a higher level, when using latent factors.

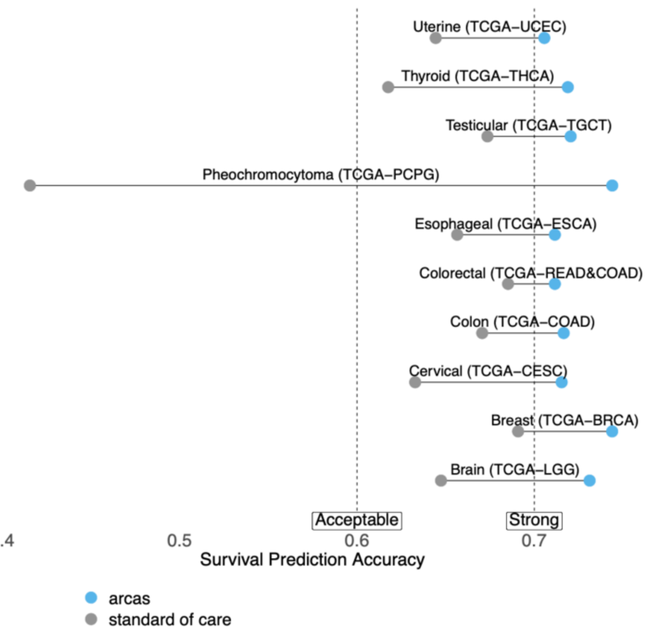

Figure 5: ARCAS platform improves survival prediction over using clinical features.

Conclusions

As sequencing prices drop, the data needed to build and run the models are getting easier to generate. In the near future, liquid biopsies and biopsies will be routinely assayed by multi-omics methods. Integrating and making sense of such datasets is the key to improve drug development and diagnostic processes. The ARCAS platform provides actionable insights from multi-omics datasets from tumor biopsies or disease models.