Identification of a molecular signature of cycling hypoxia for breast cancer prognosis: a large scale data analysis approach

Keywords : Machine learning; Feature selection; High throughput genomic data; Cancer prognosis.

Pierre Dupont, Samuel Branders

Abstract – Machine learning methods and original statistical validation protocols were developed to define a novel prognosis model for primary breast cancer. This prognosis model could help clinicians to better treat patients suffering from breast cancer. It outperforms significantly state-of-the-art prognosis kits and nicely complements clinico-pathologic criteria currently used to orient treatments. The molecular signature at the core of this result is modeling the phenotype of tumors under cycling hypoxia, i.e. the cyclic lack of oxygen of a growing tumor.

The more a tumor grows the less oxygen it receives as this source of energy gets exhausted in its surrounding environment. Tumors go through various processes to compensate for this lack of oxygen (known as hypoxia) including vascularization: the growing of new blood vessels. Their level of oxygen is then back to normal (normoxia) before a further growth and oxygen deprival phase. The more a tumor is exposed to varying levels of oxygen (i.e. cycling hypoxia), the more it is expected to develop mechanisms to survive under this stress. The phenotype of tumors conditioned by cycling hypoxia is thus expected to be related to their ability to grow and, eventually, to characterize their aggressiveness. The above hypothesis has proved successful to form the basis of a powerful prognosis tool for breast cancer [1, 2].

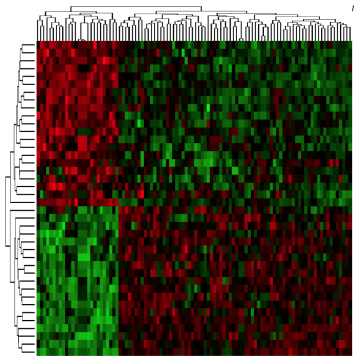

Figure 1: Gene expression signature identified from 20 tumor cell lines under normoxia versus cycling hypoxia.

At first glance, these scientific questions purely belong to cancer research but they also raise computational and statistical challenges for which the expertise of data scientists is required. Gene expression measurements on cell line data were performed on DNA microarrays from the Affymetrix GeneChip Human Genome 1.0 ST platform. This technology interrogates 28,869 well-annotated genes with 764,885 distinct probes. Twenty cell lines under 2 conditions (normoxia versus hypoxia) represent more than 30 million expression values at the probe level. Identifying an informative signature (see Figure 1) from such a large data collection requires a careful data analysis pipeline. From a statistical viewpoint, this amounts to find a needle in a haystack since it includes orders of magnitude more dimensions (the gene expression values) than samples. Validation on breast cancer data raises additional challenges to transfer a signature estimated on cell lines to real tumor tissues. Dedicated machine learning methods [3, 4], implemented in grid or cloud computing architectures, are also needed to limit the risk of overfitting which would lead to define a model looking good on an initial collection of patients but which would generalize poorly to new patients.

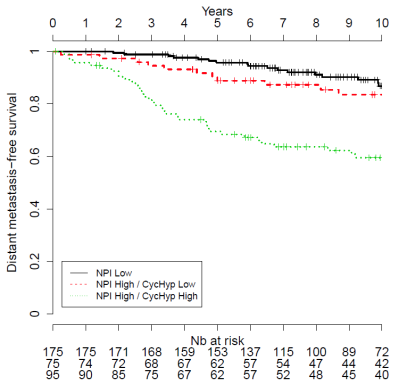

Figure 2: The proposed prognosis model based on a cycling hypoxia signature is able to detect false positives from the Nottingham Pronostic Index (NPI) currently used in clinics to orient treatments of primary breast cancer.

The cycling hypoxia prognosis model has shown to significantly outperforms state-of-the-art kits for breast cancer prognosis. This is especially true for a sub-population of ER+, HER2-, node negative patients. This result is particularly interesting since the prognosis is highly uncertain for this sub-population when considering only clinico-pathologic criteria, which are commonly used in clinics. In other words, the identified signature opens new opportunities for specific treatment guidance on top of clinical criteria routinely used.

Ongoing work aims at transferring these promising results reported for breast cancer to the prognosis of colorectal cancer. This is motivated by the fact that the modeling of cycling hypoxia is expected to be also relevant in this case and by the high prevalence of colorectal cancer. This work results from a joined research project between ICTEAM researchers and the team of Prof. O. Feron at IREC, the UCL Institute for Clinical and Experimental Research.

References

[1] R. Boidot, S. Branders, T. Helleputte, L. Illan Rubio, P. Dupont and O. Feron, "A generic cycling hypoxia-derived prognostic gene signature: application to breast cancer profiling", Oncotarget, Vol. 5, No. 16, pp. 6947-6963, July 2014.

[2] O. Feron, R. Boidot, S. Branders, P. Dupont, T. Helleputte, "Signature of cycling hypoxia and use thereof for the prognosis of cancer", WO/2015/015000, Patent Publication Date: 05/02/2015.

[3] S. Branders, "Regression, classification and feature selection from survival data : modeling of hypoxia conditions for cancer prognosis", UCL PhD. Thesis, October 2015.

[4] S. Branders and P. Dupont, "A balanced hazard ratio for risk group evaluation from survival data", Statistics in Medicine, Vol. 34, Issue 17, pp. 2528-2543, April 20, 2015.