Application of Machine Learning in Cancer Prediction & Prognosis

<http://journals.sagepub.com/doi/abs/10.1177/117693510600200030>

Growing dependences on protein biomarkers, microarray data

Strong bias towards applications in prostate in breast cancer

Reliance on old technologies such as ANN

Main application: using machine learning methods to identify, classify, detect, or distinguish tumors and other malignancies

Diagnosis, detection

Prediction, prognosis

1) the prediction of cancer susceptibility (i.e. risk assessment); 2) the prediction of cancer recurrence and 3) the prediction of cancer survivability.

Typically histological (cell-based), clinical (patient-based) and demographic (populationbased) information must all be carefully integrated by the attending physician to come up with a reasonable prognosis.

Ideally what is needed is some very specific molecular details about either the tumor or the patient’s own genetic make-up (Colozza et al. 2005).

Molecular biomarkers, such as somatic mutations in certain genes (p53, BRCA1, BRCA2),

the appearance or expression of certain tumor proteins (MUC1, HER2, PSA)

chemical environment of the tumor (anoxic, hypoxic)

combinations or patterns of multiple molecular biomarkers have been found to be even more predictive than single component tests or readouts (Savage and Gascoyne 2004; Petricoin and Liotta 2004; Duffy 2005; Vendrell et al. 2005)

If these molecular patterns are combined with macro-scale clinical data (tumor type, hereditary aspects, risk factors), the robustness and accuracy of cancer prognoses and predictions improves even more.

We also found that almost all predictions are made using just four types of input data: genomic data (SNPs, mutations, microarrays), proteomic data (specific protein biomarkers, 2D gel data, mass spectral analyses), clinical data (histology, tumor staging, tumor size, age, weight, risk behavior, etc.) or combinations of these three.

<https://www.sciencedirect.com/science/article/pii/S2001037014000464>

lack of external validation or testing regarding the predictive performance of their models.

However, these methods suffer from low sensitivity regarding their use in screening at early stages and their difficulty to discriminate benign from malignant tumors.

studies with larger data samples and more adequate validation are needed.

Pre-processing techniques: mong these techniques some of the most important approaches include (i) dimensionality reduction (ii) feature selection and (iii) feature extraction.

 Several studies in the literature discuss the phenomenon of lack of agreement between the predictive gene lists discovered by different groups, the need of thousands of samples in order to achieve the desired outcomes, the lack of biological interpretation of predictive signatures and the dangers of information leak recorded in published studies

The quantitative metrics of accuracy and AUC are used for assessing the overall performance of a classifier.

On the contrary, AUC is a measure of the model's performance which is based on the ROC curve that plots the tradeoffs between sensitivity and 1-specificity

Accuracy is a measure related to the total number of correct predictions

**Methods of splitting data:**

plitting the initial labeled data into subsets are: (i) Holdout Method-  data samples are partitioned into two separate sets, namely the training and the test sets.

, (ii) Random Sampling -  Holdout method is repeated several times, choosing the training and test instances randomly

, (iii) Cross-Validation - each sample is used the same number of times for training and only once for testing.

and (iv) Bootstrap -  samples are separated with replacement into training and test sets, i.e. they are placed again into the entire data set after they have been chosen for training.

Models:

 list of ML methods including (i) ANNs - Their generic layered structure proves to be time-consuming while it can lead to very poor performance. Additionally, this specific technique is characterized as a “black-box” technology.

, (ii) DTs – tree-structured classification scheme where the nodes represent the input variables and the leaves correspond to decision outcomes

, (iii) SVMs – SVMs are a more recent approach of ML methods applied in the field of cancer prediction/prognosis. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples.

and (iv) BNs - produce probability estimations rather than predictions

When dealing with cancer prognosis/prediction one is concerned with three predictive tasks:

(i) the prediction of cancer susceptibility (risk assessment),

(ii) the prediction of cancer recurrence/local control and

(iii) the prediction of cancer survival.

In the first two cases one is trying to find

(i) the likelihood of developing a type of cancer and

(ii) the likelihood of redeveloping a type of cancer after complete or partial remission.

In the last case, the prediction of a survival outcome such as disease-specific or overall survival after cancer diagnosis or treatment is the main objective.

The prediction of cancer outcome usually refers to the cases of

(i) life expectancy, (ii) survivability, (iii) progression and (iv) treatment sensitivity

Molecular biomarkers, cellular parameters as well as the expression of certain genes have been proven as very informative indicators for cancer prediction.

PubMed, Scopus

[genomic, clinical](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/clinical-genomics), histological, imaging, demographic, epidemiological data or combination of these

Papers that focus on the prediction of cancer development by means of conventional statistical methods (e.g. chi-square, Cox regression) were excluded as were papers that use techniques for tumor classification or identification of predictive factors

<https://www.nature.com/articles/nature03702>

Paper compares miRNA profiling accuracies

miRNA profiling

miRNA only ~21 neuleotides

methods of profiling:

1. glass-slide microarray – problem: cross-hybridaization
2. bead-based profiling

hierarchical clustering of the samples using miRNA profiles paralleled the developmental origins of the tissues. (related to primary site)

Furthermore, the miRNAs partitioned tumours within a single lineage. For example, we examined the miRNA profiles of 73 bone marrow samples obtained from patients with acute lymphoblastic leukaemia (related to mutation type: 1) BCR/ABL, TEL/AML, 2) T-cell, 3) MLL )

miRNA expression patterns encode the developmental history of human cancers.

Poorly differentiated tumor has lower levels of miRNA expression

We have previously reported that there are no robust mRNA markers that show consistent differential expression between tumours and normal tissues of different lineages

Our observation that miRNA expression seems globally higher in normal tissues compared with tumours led us to the hypothesis that global miRNA expression reflects the state of cellular differentiation.

In addition, unlike mRNAs, miRNAs remain largely intact in routinely collected, formalin-fixed, paraffin-embedded clinical tissues

The mechanism by which miRNAs are under-expressed in cancer remains unknown. We did not observe substantial decreases in the mRNAs encoding components of the miRNA processing machinery (Dicer, Drosha, Argonaute2 or DGCR8 (ref. [24](https://www.nature.com/articles/nature03702#ref24)); see [Supplementary Information](https://www.nature.com/articles/nature03702#s1)), but clearly other mechanisms of miRNA regulation are possible.

We speculate that abnormalities in miRNA expression might similarly contribute to the generation or maintenance of ‘cancer stem cells’, recently proposed to be responsible for cancerous growth in both leukaemias and solid tumours

<https://cancergenome.nih.gov/abouttcga/aboutdata/datalevelstypes>

Mutation

DNA (whole exome sequence)

Somatic mutation

Protected mutation

Idea:

Combine gene expression, isoform expression, miRNA expression quantification

<https://www.quora.com/Is-it-true-that-we-all-have-cancerous-cells-in-our-body>

Find genes related to DNA repair, cause apoptosis

controls the signaling process or the operational processes for apoptosis.

Cancer is bad, no doubt. We ALL generate cells that could be labeled “cancer,” but the vast majority never develop into anything important because our bodies have in place such a cool set of systems to prevent these cancers from progressing.