Supervised Methods for Biomarker Detection from Microarray Experiments

Serra, Angela

Springer, UK
2022

Serra , A , Cattelani , L , Fratello , M , Fortino , V , Kinaret , P A S & Greco , D 2022 , Supervised Methods for Biomarker Detection from Microarray Experiments . in G Agapito (ed.) , Microarray Data Analysis . Methods in Molecular Biology , vol. 2401 , Springer, UK , New York, NY , pp. 101-120 . https://doi.org/10.1007/978-1-0716-1839-4_8

http://hdl.handle.net/10138/355777
https://doi.org/10.1007/978-1-0716-1839-4_8

acceptedVersion

Angela Serra[1,2], Luca Cattelani[1,2], Michele Fratello[1,2], Vittorio Fortino[3], Pia Kinaret[4], Dario Greco[1,2,4,5,*]

[1] Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland.

[2] BioMediTech Institute, Tampere University, Tampere, Finland

[3] Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland.

[4] Institute of Biotechnology, University of Helsinki, Helsinki, Finland

[5] Finnish Center for Alternative Methods (FICAM), Finland.

* Correspondence: dario.greco@tuni.fi;

Running head: biomarker detection from microarray experiments

Supervised methods for biomarker detection from microarray experiments

Abstract

Biomarkers are valuable indicators of the state of a biological system. Microarray technology has been extensively used to identify biomarkers and build computational predictive models for disease prognosis, drug sensitivity and toxicity evaluations. Activation biomarkers can be used to understand the underlying signalling cascades, mechanisms of action and biological crosstalk. Biomarker detection from microarray data requires several considerations both from the biological and computational points of view. In this chapter, we describe the main methodology used in biomarkers discovery and predictive modelling and we address some of the related challenges. Moreover, we discuss biomarker validation and give some insights into multi-omics strategies for biomarker detection.

Keywords (5-10): microarray; biomarker; classifier; feature selection; validation metrics; data unbalancing; model selection; hyper-parameters estimation; biological validation; multi-omics

Introduction

The biological state of a system can be defined in terms of activated, deactivated or altered indicators. These known (bio)markers can be measured from the molecular, biochemical, cellular, physiological, pathological or behavioral state in the biological system due to a changed condition, such as drug/chemical exposure or disease stage ***(1–3).*** Considering the high heterogeneity between individuals, different disease stages, and the complexity of biological systems, usually a

single marker does not provide enough value or power for comprehensive conclusions and predictions. To determine a descriptive panel of meaningful biomarkers, advanced, high-throughput methods are required. With microarrays, a substantial panel of possible molecular biomarkers - genes with expressional alterations can be determined simultaneously. Microarrays provide an in-depth method for the identification of specific gene signatures and patterns with high predictive value. The obtained information can also be used to build a detailed and inclusive understanding of the underlying biological state as well as to classify and predict the disease onset or chemical hazard. In disease diagnostics, some biomarkers are easily measured with modern laboratory techniques. For example, specific antibodies suggest a specific treatment against a pathogen while different lung function tests such as spirometry, fractional exhaled nitric oxide or peak flow are indicative markers of asthma development. With no predictive power, these markers are used to determine treatments for already existing diseases. Instead, microarrays and high throughput techniques have enabled a shift from the traditional medical and therapeutic approaches towards predictive and precision techniques, utilising sophisticated computational methodologies and algorithms. Simultaneously, the large amount of data facilitates the use of biomarkers in data modelling, allowing a more detailed understanding of the chemical and drug sensitivity and toxicity, such as dose- and time-dependency for the risk and hazard assessment.

In biomarker analysis from microarray data, two main computer-aided tasks can be performed namely biomarker discovery and development of predictive modelling ***(4–7)***. The former refers to the identification of the smallest, most accurate and reliable set of predictive biomarkers for a particular endpoint. The latter refers to the development of a computational model that, using the subset of identified biomarkers, can learn a function that connects their expression values to a phenotypic outcome. However, microarray data pose some computational challenges that need to

be addressed for biomarker discovery and predictive modelling development to succeed ***(7–11)***. Since the number of biological samples is usually limited compared to the number of measured bio-molecules, such as genes, microarray experimental data can contain noisy information. However, exploring all the possible subsets of tens of thousands of biomolecules in the microarray experiment is computationally infeasible. This problem is tackled by feature selection methods that reduce the risk of overfitting and the computational burden. These methodologies include simple univariate and multivariate statistical analysis or more complex machine learning-based algorithms ***(12)***.

Predictive models can be categorized as classification or regression methods depending on whether the predicted variable is categorical or continuous ***(12, 13)***. Simpler models, that directly use the values of few biomarkers to perform a prediction, are easy to interpret and help enlarge the understanding of the biological process under study. However, they do not always ensure the highest predictive capability unlike more complex models that use many biomarkers or non-linear combinations of them. These models can lead to better predictive performances, however, deriving a biological interpretation is more difficult. To reach the highest levels of predictive power, these algorithms may require extensive tuning of their input parameters. Moreover, since the experimental data are highly heterogeneous, it is important to evaluate these models on external independent datasets that are often unavailable.

In this chapter, we will describe the most common feature selection and predictive methods for biomarker discovery. Moreover, we will discuss the challenges related to the use of machine learning methods such as model selection, parameter tuning, and reproducibility. We will shortly introduce the issue related to data unbalancing and we will discuss the biological validation of the

results. Culminating in a short overview of the multi-omics methodologies available for biomarker detection.

Feature selection based approaches for biomarker discovery

Biomarker discovery methods from microarray data aim to identify the smallest, most accurate and reliable set of predictive biomolecules. This task is usually performed by applying feature selection algorithms to microarray data. Feature selection is the process by which a subset of relevant biomarkers is selected to construct accurate predictive models. In the context of supervised learning, feature selection techniques can be divided into three main categories: filter, wrapper and embedded (Figure 1) ***(14)***.

[Figure 1 near here]

Filter methods evaluate the relevance of the features by only looking at the intrinsic properties of the data, independently from the selected classifier. These methods compute feature relevance scores and only top-ranked features are presented to the classifier. Because of the high dimensionality of omics datasets, fast, univariate filters have been widely applied. For gene expression data, the simplest heuristic is to rank the genes according to their deregulation between the treated and the control samples. The main assumption is that genes with the strongest expression change at the top of the ranks represent the key drivers of the disease stage or response, and thus, are chosen as candidate biomarkers for further validation.

Multiple parametric and non parametric methodologies are widely applied such as the two sample *t*-test and ANOVA or the Wilcoxon rank-sum test and the information gain ***(15)***. However, these univariate methodologies do not take into account the feature dependencies and may lead to less

accurate classification. To this end, multivariate filter methodologies have been suggested ranging from simple bivariate interactions towards advanced solutions exploring higher-order interactions. The correlation-based feature selection ***(16)***, the ReliefF ***(17)***, and the Minimum Redundancy-Maximum Relevance (MRMR) ***(18)*** are examples of solid multivariate filter procedures, highlighting the advantage of using multivariate methods over univariate procedures in the gene expression domain.

Wrapper methods mix the feature selection step together with the model parameter search. In this scenario, subsets of features are evaluated by training and testing a specific classification algorithm. The wrapping methods can be categorized into deterministic, which try to explore all the possible subsets of features, and randomized. The forward- and backward- sequential selection and SVM-RFE are examples of deterministic wrapper methods ***(19–21)***. These algorithms did not receive a lot of attention in the omics data analysis literature since exploring all the feature subspace is a limitation when tens of thousands of features are considered. On the other hand, randomized approaches such as particle swarm optimization and genetic algorithms have been applied in the omics data analysis ***(22, 23)***. For example, the GARBO method, based on a genetic algorithm for biomarker discovery and feature set optimization, has recently been proposed. GARBO identifies the smallest and most robust set of biomarkers with the best predictive performances from a single-omics data layer ***(24)***.

In the embedded methods, the search for the optimal set of features is built into the classifier construction, making them specific to the learning algorithm. These approaches are less computationally intensive than the wrapper methods since they do not search for all possible subsets of features. Examples of embedded feature selection methods are feature importance

derived from decision trees and random forests (RF) ***(25–27)***, and regularization based methods, such as Ridge, LASSO and ElasticNet ***(28–30)***.

Predictive modelling

The task of prediction refers to the development of a computational model that, using the subset of identified biomarkers, is able to learn a function that connects their expression values to a phenotypic outcome (*e.g.*, toxicity assessment or disease severity) ***(12, 31)***. The task of predicting discrete values is known as classification, whereas when the outcome variable is continuous we talk about regression. An example of classification is the task of discriminating between toxic or non-toxic compounds. On the other hand, an example of regression would be the prediction of a drug sensitivity score.

[Figure 2 near here]

Classification based predictive modelling

Classification is the problem of assigning samples, represented by vectors of features, to a specific class between a set of possible ones. A classifier is a function that maps a sample to a class. In the context of biomarker detection, the classes are the outcomes of interest (*e.g.*, toxic/non-toxic, drug resistant/non-resistant), while the features are potential biomarkers (Figure 2).

Some properties of classifiers are particularly useful in the context of biomarker detection. One is parsimony in the number of features used, if a classifier uses only the strictly necessary number of features, it is easier to understand and to apply in practice, and less prone to overfit with the inclusion of spurious features that are not real biomarkers. A common distinction is between

"white-box" and "black-box" models. White-box models allow for an easy understanding of the underlying algorithm that leads from the features to the assigned class, as opposed to black-box models that are difficult to grasp, due to a high number of parameters and/or complex non-linear interactions. White-box algorithms are generally more appreciated since they can lead to better insights into how biomarkers and biological outcome are related.

[Figure 3 near here]

Machine learning and artificial intelligence techniques are pervasive in biological studies ***(13)***, like cancer drug resistance ***(32)***, or chemical toxicity assessment ***(12)***. Supervised methods allow to train a model, a classifier in this case, starting from training data. Training data includes feature vectors and class labels. A trained model can then be applied to new labeled data for validation or to unlabeled data for predictions. We will briefly describe four of the most well known classification methods: logistic regression (LR) (Figure 3A), support vector machine (SVM) (Figure 3B), random forest (RF) (Figure 3C), and artificial neural network (ANN) (Figure 3D), together with some examples of application to biomarker detection using microarray data.

A LR is composed of a standard logistic function applied to the result of a linear function, with the parameters of the linear function that are learned on the training data, typically with the maximum likelihood method. Park et al. ***(33)*** proposed a novel penalization method that incorporates a measurement of the significance of genes to LASSO-type regularization, and used it to classify cell lines as drug sensitive or resistant, and identify biomarkers, on the Sanger dataset from the Cancer Genome Project (www.cancerrxgene.org).

SVM separates two classes of training samples seen as points in an n-dimensional space, where n is the number of features of the samples, by an *n-1*-dimensional hyperplane, so that a new unlabelled sample is classified according to the side on which it is placed with respect to the hyperplane. If the training samples cannot be separated by a hyperplane, it is possible to apply a non-linear transformation to the feature vectors so that the resulting vectors are now separable ***(34)***. Zheng et al. ***(35)*** applied two kinds of SVM and logistic regression to Serum miRNA expression profiles from *52* Esophageal Squamous Cell Carcinoma patients, and identified miR-16-5p, miR-451a, and miR-574-5p as biomarkers for the diagnosis of the disease. There was substantial concordance on the choice of features, while SVMs showed slightly better predictive performance than logistic regression.

Classification trees in RF method, are tree structures representing decision processes where starting from the root, at each branch, an evaluation is made on the input features that assigns the process to one of the following branches. This is repeated until a leaf is reached, each leaf representing a decision on the class of the sample. RF for classification are grown by training a number of classification trees on different extractions of the training samples, performed with the bagging method ***(25–27)***. Su et al. ***(36)*** used data from the HT-HGU133A Affymetrix whole genome array belonging to the Cancer Cell Line Encyclopedia and the Genomics of Drug Sensitivity in Cancer database. They compared SVM, Deep Forest ***(37)***, and a new deep forest-based algorithm in classifying drug response into “sensitive” or “resistant”. Results showing slightly better predictions of the deep forest algorithms and a substantial equality between these two.

Artificial neural networks (ANNs) are based on networks of units in which each non-input unit integrates signals from its predecessors, typically with a dot product operation, and then applies a

nonlinear function, *e.g.*, a sigmoid function. In feed forward ANNs the units are arranged in a directed acyclic graph, while in recurrent neural networks the graph is cyclic. The input units are fed with the input features, internal units process signals from other units, and the output units return the output classification. Each non-input processing unit has a set of parameters that are learned during training, typically with a back-propagation algorithm ***(38)***. Wang et al. ***(39)*** used Affymetrix GeneChip Rat Genome 230 2.0 Array *in vivo* liver data from DrugMatrix and Open TG-GATEs, to train and validate SVM, RF, and single and multi-task deep neural networks (DNN) on the tasks of predicting biliary hyperplasia, fibrosis and necrosis, in order to compare the accuracy of the models. Single-task DNN and SVM outperformed RF and multi-task DNN for the three endpoints. The two best models were further compared on another dataset (Gene Expression Omnibus accession number, GSE70559) where Single-task DNN outperformed SVM.

Regression based predictive modelling

Regression is a supervised learning methodology that estimates the relationship or function between the features and a continuous variable (Figure 4). Regression methods for biomarker detection from microarray data have been applied to predict important quantities such as the toxicity level of a compound ***(40–42)***, the drug sensitivity tumor cell lines ***(43, 44)***, and patients survival ***(45, 46)***. Moreover, regression based methods are extensively used in toxicogenomics to identify dose-responsive genes. Under the hypothesis that dose-responsive genes are altered as a direct consequence of the exposure, they can be prioritized as candidate biomarkers for the biological question under study ***(12, 47–50)***.

The most common regression algorithm is the linear regression where all the features (e.g., genes) are linearly combined to predict an outcome variable. In case of high dimensional data, such as the

ones coming from microarray experiments, the linear regression method can be combined with regularization methods which are able to estimate the contribution of the different variables to the overall prediction problem. Examples of these regularization methods are LASSO ***(29)*** and ElasticNet ***(51)*** regularization. When these techniques are used, the biomarker discovery and modelling steps are embedded into the predictive modelling resulting in smaller sets of biomarkers. ElasticNet is a linear regression with a hybrid regularization term combining LASSO and Ridge regularizations ***(51)***.

An example of application of regression models for drug sensitivity prediction from cancer cell data is the work of Jang et al. ***(52)***, where the authors compared multiple models for the analysis of pharmacogenomic datasets in search of biomarkers for continuous drug sensitivity scores. Their results suggested that ElasticNet or Ridge regression methods working on the whole set of genomic features, in particular those coming from gene expression profiles, yield the most accurate predictions. Similarly, Ding et al. ***(43)*** applied the ElasticNet regression to generate logistic models for drug sensitivity prediction in the Cancer Cell Line Encyclopedia and Genomics of Drug Sensitivity in Cancer project datasets.

[Figure 4 near here]

Validation Metrics

When training a predictive model, it is usually assumed that the dataset collected is representative of the underlying data distribution, *i.e.*, new, unseen data should "look like" the data collected for training. The objective is to train a model to learn the distribution of the data reasonably enough to be able to generalize appropriately to new data samples and to avoid overfitting. The easiest

approach to evaluate the generalization capabilities of a trained model is to split the dataset into a training set and a test set. Once trained, the model's generalization capabilities are estimated with a variety of measures ***(53, 54)***.

Accuracy Measures in classification

In the case of a binary classification problem, such as the discrimination of toxic *vs.* non toxic chemicals, the goodness of the model can be evaluated by computing accuracy measures from a confusion matrix. For example, out of *100* chemicals tested, *30* are toxic and *70* are non toxic. This scenario is summarized in Table 1.

The confusion matrix shows the following values:

- true positives (*TP*): number of samples from the positive class (Toxic) correctly classified as such (~Toxic);
- false negatives (*FN*): number of samples from the positive class (Toxic) classified as negative samples (~Non Toxic);
- false positives (*FP*): number of samples from the negative class (Non Toxic) classified as positive samples (~Toxic);
- true negatives (*TN*): number of samples from the negative class (Non Toxic) correctly classified as such (~Non Toxic);

With these quantities, different predictive metrics can be defined, such as:

Recall or sensitivity, hit rate or true positive rate: defined as *TP / (TP + FN)*, which corresponds to the portion of positive data points which are correctly considered as positive, with respect to all the positive data points. High values imply few false negatives; In our example this would be *25 / (25 + 5) ≈ 0,83*.

Specificity, selectivity or true negative rate: defined as $TN / (TN + FP)$, that measures the proportion of negative data points that are correctly identified. High specificity indicates the presence of a few false positives. In our example this would be $60 / (60 + 10) \approx 0,86$.

Precision: defined as $TP / (TP + FP)$, whose high values imply few false positives; In our example this would be $25 / (25 + 10) \approx 0,71$.

Accuracy: defined as $(TP + TN) / (TP + TN + FN + FP)$, whose values define the proportion of correctly classified samples compared to the samples in the dataset; in our example this would be $(25 + 60) / 100 = 0,85$.

***F1*-score (or *F*-score or *F*-measure):** defined as $2 \times TP / (2 \times TP + FP + FN)$, it is the harmonic mean of precision and recall. In our example this would be $2 \times 25 / (2 \times 25 + 10 + 5) \approx 0,67$.

According to the context, high rates of false negative or false positive predictions can have different implications such as using a compound predicted to be non toxic as a treatment when it is actually toxic (false negative) or *vice-versa* not assigning a compound as a treatment because it is predicted to be toxic, when it is actually non toxic (false positive). The *F1*-score was defined to find a balance between the precision and recall metrics.

All of these metrics values range between *0* and *1*. Good performances are achieved for values as close as possible to *1*. In the case of a binary, the proportion of the most represented class (*0.5* in case of balanced classes) can be used as a threshold for chance level, meaning the accuracy that a classifier would get, if it randomly assigns the majority class instead of using a model. Models are said to have predictive power when they perform better than the chance level.

Accuracy and *F1*-score are widely used to evaluate classification models, however, in cases where the data is heavily unbalanced, these measures alone are inappropriate ***(55)***.

Data unbalancing

Usually, the amount of samples in one of the classes is significantly outnumbered by the samples of the opposite class. For example, in the case of toxic vs. non toxic drugs, the number of toxic compounds in the dataset is often much lower than the non toxic ones ***(56–58)***. The more a dataset is unbalanced, the less reliable some of the previous metrics become. For example, in the case of a dataset with *100* drugs with *10* toxic and *90* non toxic drugs, a classifier that always predicts the drugs to be non toxic would have an accuracy of *90%* even though it completely misclassifies the toxic category. This is because, during training, it's easier to learn the negative class by increasing the number of false negatives. In these cases some particular strategies need to be applied in order to perform a good evaluation of the model.

To compensate for imbalances, the samples can be made more or less relevant by weighting each class with the inverse of the corresponding class proportion. A more elaborate approach consists in resampling parts of the dataset: the majority class can be down-sampled (*i.e.* randomly discard a number of samples), the minority class can be over-sampled, or both. Over-sampling can be as simple as randomly adding duplicate samples, or it can be a generative scheme that creates new synthetic samples combining the actual samples such as ROSE ***(59, 60)***, SMOTE or its variants ***(61, 62)***.

In conjunction to these approaches, model evaluation should be performed using a metric that takes into account the proportions of each possible outcome (*TF*, *TN*, *FP*, *FN*) such as the Matthews Correlation Coefficient ***(55, 63)***, defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

The *MCC* varies between *-1* and *1*, with *1* being the best score, *i.e.* perfect classification. When the *MCC* is *0*, the classifier is equivalent to a random guess. Finally, when the *MCC* equals *-1*, the classifier predicts each sample with the opposite label.

Goodness of fit measures in regression

A model is considered accurate when the difference between the real and predicted values (*i.e.*, the error) is as small as possible. Given a vector of real values y and predicted values $\tilde{y}$, the most commonly adopted error measure is the mean absolute error (*MAE*) (Figure 5A) defined as

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|y_j - \tilde{y_j}|$$

However, the *MAE* does not give any information on the direction of the error, for example, if a value is under or over predicted. Another commonly used metric is the mean squared error (*MSE*) (Figure 5A) defined as

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - \tilde{y_j})^2$$

The *MSE* is more susceptible to outliers in the dataset since *MSE* evaluates the square of the error, while the *MAE* is less affected, so this is a point to take into account while choosing a proper error measure. Other variants of *MAE* and *MSE* are the root mean squared error (*RMSE*), a scale-independent alternative of *MAE* called *RAE*, and the relative squared error (*RSE*) ***(53)***.

Even though the model performs better when the errors are as low as possible, these metrics are not so easy to interpret such as the comparison with chance level of the binary classification problem. To this end the R^2 metric can be used to compare the model performance against a

baseline level, which is the mean value of the variable to predict (Figure 5B). Given a vector of real values y and predicted values $\tilde{y}$, the R^2 metric is defined as follows:

$$R^2 = \frac{MSE(model)}{MSE(baseline)} = \frac{\sum_{i=n}^{N}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{N}(\bar{y} - \tilde{y}_i)^2}$$

where $\bar{y}$ is the mean value y. In other words, the predictive performances of the model are compared with those of a model that always predict the mean value. R^2 values can range in *0-1* with values close to *1* being the best as possible, while a model performing equal to the baseline would give value of *0*. However, this measure does not take into account the fact that the more features (biomarkers) are used by the model, the closer the value will be to *1*. Thus, an adjusted formula of the R^2 can be used to penalize models which use a lot of features compared to the number of samples. The adjusted R^2 is defined as

$$R^2_{adj} = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right]$$

where *k* is the number of features and *n* is the number of samples. The adjusted R^2_{adj} measure can be used to evaluate the feature importance: when adding a relevant feature to the model the adjusted R^2_{adj} increases. If a new feature is added and the value does not increase, it means that the added feature is not relevant.

[Figure 5 near here]

Model selection and Hyper-parameter optimization

In addition to model parameters that are learned during training, most models also have a set of hyper-parameters that need to be tuned to achieve optimal performances, like the number of trees in RF, the architecture of a neural network or the value of the regularization parameter of the LASSO method.

These hyper-parameters cannot be inferred directly from data like other training parameters, and need to be estimated by means of an explicit search. This procedure usually requires the dataset to be split in to at least three disjoint subsets: the training and validation sets that are used for model fitting and hyper-parameter selection, and the test set, that is only used for the final evaluation and never to further tune the models ***(64)***. A common rule of thumb is to use roughly *65%* of the samples for training, *15%* of the samples for validation and *20%* of the samples for testing. However, splitting a dataset reduces the available data; a more data-efficient approach is *k*-fold cross-validation, in which the dataset is randomly split into *k* subsets of approximately the same size, then iteratively, one of the *k* subsets is used as a validation set and the remaining *k-1* subsets as training. The cross-validated estimate is then the average across the *k* runs. Also in this case, there is no set rule to choose *k,* if not as a trade-off between the stability of training and reliability in validation, common choices for *k* are *5* or *10*. The limit case where *k* is equal to the number of samples is called leave-one-out cross-validation.

External Validation of Biomarkers

It is estimated that the current number of candidate biomarker panels based on omics data is over one million. However, only few of them have been successfully translated into clinically useful tests leading to the so-called biomarker innovation gap ***(65)***. A major factor contributing to this gap is the challenge of assessing whether the body of evidence of omics-informed biomarkers is sufficiently reliable or not ***(66)***. A way to assess and increase the reliability of omics-based biomarkers before clinical testing, is to verify their prediction performances on external verification data sets, which can be retrieved from public repositories. External validation is necessary to reduce model instability and data overfitting ***(67)***. However, automatizing the search

and the re-use of publicly available data for biomarker verification and refinement is a laborious task.

External validation datasets can be retrieved from public repositories that archives and freely distributes omics datasets, such as ArrayExpress ***(68)***, Gene Expression Omnibus (GEO) ***(69)***, GenomeRNAi ***(70)*** and dbGAP ***(71)***. These databases include thousands of different omics datasets, and often researchers struggle in discovering 'similar' omics datasets. During the past few years, different platform search engines have been proposed to find and link existing omics datasets. Table 2 includes a list of published search engines that researchers can use to link omics studies with a similar experimental setup (*e.g.*, same disease, same tissue, similar omics technology, similar clinical phenotype, *etc.*). These search engines provide application-programming interfaces (APIs) to query and access their data programmatically.

Biological Validation

In the optimal case, a biomarker(s) leads to an accurate and precise prediction of a biological endpoint. Due to natural heterogeneity of biological samples and the unavoidable technical biases, the biomarker detection and/or predictions are not simple objectives. In order to confirm the validity of the detected/predicted microarray biomarkers, a measure of the real abundance and the statistically significant effect is probably needed. These measures also require repeatability, meaning that the outcome is detected from repeated but distinct experiments. To measure the real, biological abundance of a gene transcript instead of the relative fold-changes obtained from microarray data, quantitative polymerase chain reaction (qPCR) technology is often used. qPCR is considered a state-of-the-art validation step to measure transcriptional activation. However, measuring the same samples with two distinct methods such as microarrays and qPCR, mainly

provides information about the possible variance between two different technologies. Thus, to understand the real biological significance, a completely new set of samples should be prepared and measured. This, however, is not often executable with biological samples. Instead of utilizing additional technical measures for validation, more emphasis should be put into the interpretation of the biological meaning behind the data to recognise other important regulatory cascades. For this, studying the upstream regulators or co-regulators from the same microarray data set can explain/confirm the expressional changes measured or predicted from the microarray data sets. Also other techniques explaining the biological events behind the data, such as immunohistochemistry, fluorescence *in situ* hybridization or chromatography can be successfully utilized for validation of the microarray data ***(72)***. Moreover, it should be noted that transcriptional change measured by microarrays, does not necessarily inform about the translational changes and the consequent protein product. Thus, to validate the existence of the actual gene product other experimental techniques might be required. Although accurate and reproducible biomarker(s) with high predictability are discovered through computational modeling and validation steps, for clinical or regulatory purposes the evaluation will be continued in terms of patient samples and or additional animal models.

Multi-omics strategies

Due to technological advances, different types of omics data have become available. Among them are gene expression, microRNA expression, copy number variation, methylation, and SNP. This allows the measurement of multiple omics data layers for the same set of samples. These multi-omics experimental data are often not highly correlated between each other, thus they provide potentially complementary information and assess different parts of the same complex biological process ***(73, 74)***.

Multiple data integration strategies to merge and analyze multi-omics data arise in a wide range of clinical, toxicogenomics, and functional genomics applications ***(9, 73, 75–77)***. Depending on the type of data integration strategy, integrative multi-omics data analysis can be classified into early, intermediate or late integration ***(13, 78, 79)***. In the early integration strategies, the multi-omics data layers are merged in a single dataset with the same samples and a number of features equal to the sum of the features of the different data layers. In the intermediate integration, the single omics layers are first individually transformed in the same space and then combined in a single dataset on which the feature selection or predictive algorithms are applied. In the late integration approaches, each algorithm is executed independently and in parallel on each omic layer, and only in the end the results of each algorithm are integrated.

Some of the classical machine learning algorithms have been adapted to the analysis of multi-omics dataset. For example, an adaptation of the min-redundancy and max-relevance (mRMR) feature selection method for multi-omics data for predicting ovarian cancer survival has been proposed ***(80)***. A classical mRMR algorithm iteratively identifies features that are of maximal relevance for the prediction task and minimally redundant (*e.g.,* not correlated) with the set of already selected features. In case of multi-omics data, the mRMR algorithm could be applied independently to each omic layer (late integration), or to a new dataset created by concatenating all the layers (early integration). However, in the first case it would be difficult to evaluate the redundancy of the features between multiple data layers. In the second case it would fail to identify differences in the relevance of features coming from different views, or features from a view could be neglected. The authors suggest a two-level approach (intermediate integration) where the mRMR algorithm is applied on each omic data layer to identify its specific relevant biomarkers.

A further step is applied on the concatenation of the features identified at the previous step to select a final set of multi-omics non redundant biomarkers.

Another example of intermediate integration is provided by ***(81)*** where a novel multi-view feature selection based on the canonical correlation analysis (CCA) statistical method was proposed. This method first identifies, by means of CCA, a common d-dimensional space among all the omics data layers and then scores and ranks the input features in this space to select the most relevant ones of each layer and combine them in a final dataset on which a classifier can be applied. In this study, the effectiveness of their methods to predict kidney renal clear cell carcinoma (KIRC) survival from copy number alteration, gene expression and reverse-phase protein array was reported.

Another example is the work of Wang et al. ***(82)*** where they used a sparse multi-view matrix factorization (sMVMF) approach for gene prioritization in gene expression data from multiple tissues. In this case, the omic feature is only one (gene expression) but the layers are represented by multiple tissue types. The authors showed the effectiveness of the sMVMF algorithm on three human tissues from the TwinsUK cohort. The sMVMF method was able to identify genes whose expression variance across multiple tissues and those that are tissue specific. This kind of approach is able to shed light on biological problems that are involved with tissue differentiation.

Conclusions

Multiple statistical and machine learning methodologies have been applied to the analysis of microarray data in search of biomarkers as indicators of the state of a biological system. In this chapter we introduced the basic concepts related to biomarker discovery and predictive modelling from microarray data, with particular attention on their related computational challenges, such as model selection and hyperparameter tuning, data unbalancing, metrics for model validations. We

also discussed the use of external data to further evaluate the predictive capabilities of the trained models and the biological validation of the identified biomarkers. Moreover, we briefly introduced multi-omics strategies for biomarkers identification. The authors hope that this short review could provide a useful compendium to bioinformatics practitioners.

Bibliography

1. Strimbu K and Tavel JA (2010) What are biomarkers? Curr Opin HIV AIDS 5:463–466
2. Gupta RC (2014) Introduction, In: Biomarkers in Toxicology, pp. 3–5 Elsevier
3. Califf RM (2018) Biomarker definitions and their applications. Exp Biol Med (Maywood) 243:213–221
4. Torres R and Judson-Torres RL (2019) Research techniques made simple: feature selection for biomarker discovery. J Invest Dermatol 139:2068-2074.e1
5. Shahrjooihaghighi A, Frigui H, Zhang X, et al (2017) An ensemble feature selection method for biomarker discovery. Proc IEEE Int Symp Signal Proc Inf Tech 2017:416–421
6. Deng X and Campagne F (2010) Introduction to the development and validation of predictive biomarker models from high-throughput data sets. Methods Mol Biol 620:435–470
7. McDermott JE, Wang J, Mitchell H, et al (2013) Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. Expert Opin Med Diagn 7:37–51
8. Piatetsky-Shapiro G and Tamayo P (2003) Microarray data mining. SIGKDD Explor Newsl 5:1
9. Deyati A, Younesi E, Hofmann-Apitius M, et al (2013) Challenges and opportunities for oncology biomarker discovery. Drug Discov Today 18:614–624
10. Kinaret PAS, Serra A, Federico A, et al (2020) Transcriptomics in toxicogenomics, part I: experimental design, technologies, publicly available data, and regulatory aspects. Nanomaterials (Basel) 10
11. Federico A, Serra A, Ha MK, et al (2020) Transcriptomics in toxicogenomics, part II: preprocessing and differential expression analysis for high quality data. Nanomaterials (Basel) 10
12. Serra A, Fratello M, Cattelani L, et al (2020) Transcriptomics in toxicogenomics, part III: data modelling for risk assessment. Nanomaterials (Basel) 10
13. Serra A, Galdi P, and Tagliaferri R (2018) Machine learning for bioinformatics and neuroimaging. WIREs Data Mining Knowl Discov e1248
14. Saeys Y, Inza I, and Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23:2507–2517

15. Hall MA and Smith LA (1998) Practical feature subset selection for machine learning.
16. Yu L and Liu H (2003) Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, In: Fawcett, T. and Mishra, N. (eds.) Proceedings, Twentieth International Conference on Machine Learning, pp. 856–863 Amer Assn for Artificial, Menlo Park, Calif
17. Kononenko I (1994) Estimating attributes: Analysis and extensions of RELIEF, In: Bergadano, F. and Raedt, L. (eds.) Machine Learning: ECML-94, pp. 171–182 Springer Berlin Heidelberg, Berlin, Heidelberg
18. Peng H, Long F, and Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238
19. Somol P, Pudil P, Novovičová J, et al (1999) Adaptive floating search methods in feature selection. Pattern Recognit Lett 20:1157–1163
20. Borboudakis G and Tsamardinos I (2019) Forward-backward selection with early dropping. 20:276–314
21. Sanz H, Valim C, Vegas E, et al (2018) SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformatics 19:432
22. Annavarapu CSR, Dara S, and Banka H (2016) Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm. EXCLI J 15:460–473
23. Chuang L-Y, Yang C-H, Li J-C, et al (2012) A hybrid BPSO-CGA approach for gene selection and classification of microarray data. J Comput Biol 19:68–82
24. Fortino V, Scala G, and Greco D (2020) Feature set optimization in biomarker discovery from genome-scale data. Bioinformatics 36:3393–3400
25. Breiman L (2001) Random forests. 45:5–32
26. Chen X and Ishwaran H (2012) Random forests for genomic data analysis. Genomics 99:323–329
27. Fratello M and Tagliaferri R (2019) Decision trees and random forests, In: Encyclopedia of bioinformatics and computational biology, pp. 374–383 Elsevier
28. Hastie T (2020) Ridge regularization: an essential concept in data science. Technometrics 1–8
29. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58:267–288
30. Zou H and Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Statistical Soc B 67:301–320
31. Larrañaga P, Calvo B, Santana R, et al (2006) Machine learning in bioinformatics. Brief Bioinformatics 7:86–112
32. Tolios A, De Las Rivas J, Hovig E, et al (2020) Computational approaches in cancer multidrug resistance research: Identification of potential biomarkers, drug targets and drug-target interactions. Drug Resist Updat 48:100662
33. Park H, Shiraishi Y, Imoto S, et al (2017) A Novel Adaptive Penalized Logistic

Regression for Uncovering Biomarker Associated with Anti-Cancer Drug Sensitivity. IEEE/ACM Trans Comput Biol Bioinform 14:771–782
34. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, et al (2020) A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 408:189–215
35. Zheng D, Ding Y, Ma Q, et al (2018) Identification of serum micrornas as novel biomarkers in esophageal squamous cell carcinoma using feature selection algorithms. Front Oncol 8:674
36. Su R, Liu X, Wei L, et al (2019) Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. Methods 166:91–102
37. Zhou Z-H and Feng J (2020) Deep Forest.
38. Abiodun OI, Jantan A, Omolara AE, et al (2018) State-of-the-art in artificial neural network applications: A survey. Heliyon 4:e00938
39. Wang H, Liu R, Schyman P, et al (2019) Deep neural network models for predicting chemically induced liver toxicity endpoints from transcriptomic responses. Front Pharmacol 10:42
40. Raies AB and Bajic VB (2016) In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip Rev Comput Mol Sci 6:147–172
41. Maunz A and Helma C (2008) Prediction of chemical toxicity with local support vector regression and activity-specific kernels. SAR QSAR Environ Res 19:413–431
42. Xu Y, Pei J, and Lai L (2017) Deep Learning Based Regression and Multiclass Models for Acute Oral Toxicity Prediction with Automatic Chemical Feature Extraction. J Chem Inf Model 57:2672–2685
43. Ding MQ, Chen L, Cooper GF, et al (2018) Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. Mol Cancer Res 16:269–278
44. Geeleher P, Cox NJ, and Huang RS (2014) Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol 15:R47
45. Wenbin Zhang, Jian Tang, and Nuo Wang (2016) Using the machine learning approach to predict patient survival from high-dimensional survival data, In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1234–1238 IEEE
46. Tong Z, Liu Y, Ma H, et al (2020) Development, validation and comparison of artificial neural network models and logistic regression models predicting survival of unresectable pancreatic cancer. Front Bioeng Biotechnol 8:196
47. Serra A, Saarimäki LA, Fratello M, et al (2020) BMDx: a graphical Shiny application to perform Benchmark Dose analysis for transcriptomics data. Bioinformatics 36:2932–2933
48. Kuo B, Francina Webster A, Thomas RS, et al (2016) BMDExpress Data Viewer - a visualization tool to analyze BMDExpress datasets. J Appl Toxicol 36:1048–1059
49. Serra A, Fratello M, Del Giudice G, et al (2020) TinderMIX: Time-dose integrated

modelling of toxicogenomics data. Gigascience 9
50. Saarimäki LA, Kinaret PAS, Scala G, et al (2020) Toxicogenomics analysis of dynamic dose-response in macrophages highlights molecular alterations relevant for multi-walled carbon nanotube-induced lung fibrosis. NanoImpact 100274
51. Friedman J, Hastie T, and Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33:1–22
52. Jang IS, Neto EC, Guinney J, et al (2014) Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput 63–74
53. Galdi P and Tagliaferri R (2019) Data mining: accuracy and error measures for classification and prediction, In: Encyclopedia of bioinformatics and computational biology, pp. 431–436 Elsevier
54. Handelman GS, Kok HK, Chandra RV, et al (2019) Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. AJR Am J Roentgenol 212:38–43
55. Chicco D (2017) Ten quick tips for machine learning in computational biology. BioData Min 10:35
56. Tharwat A, Moemen YS, and Hassanien AE (2016) A predictive model for toxicity effects assessment of biotransformed hepatic drugs using iterative sampling method. Sci Rep 6:38660
57. Tharwat A, Moemen YS, and Hassanien AE (2017) Classification of toxicity effects of biotransformed hepatic drugs using whale optimized support vector machines. J Biomed Inform 68:132–149
58. Eitrich T, Kless A, Druska C, et al (2007) Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. J Chem Inf Model 47:92–103
59. Lunardon N, Menardi G, and Torelli N (2014) ROSE: a package for binary imbalanced learning. R J 6:79
60. Menardi G and Torelli N (2014) Training and assessing classification rules with imbalanced data. Data Min Knowl Discov 28:92–122
61. Chawla NV, Bowyer KW, Hall LO, et al (2002) SMOTE: Synthetic Minority Over-sampling Technique. jair 16:321–357
62. Kovács G (2019) An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Appl Soft Comput 83:105662
63. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451
64. Hastie T, Tibshirani R, and Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. The elements of statistical learning: data mining, inference, and prediction
65. Gool AJ van, Bietrix F, Caldenhoven E, et al (2017) Bridging the translational innovation gap through good biomarker practice. Nat Rev Drug Discov 16:587–588

66. McShane LM, Cavenagh MM, Lively TG, et al (2013) Criteria for the use of omics-based predictors in clinical trials. Nature 502:317–320
67. Taylor JMG, Ankerst DP, and Andridge RR (2008) Validation of biomarker-based risk prediction models. Clin Cancer Res 14:5977–5983
68. Athar A, Füllgrabe A, George N, et al (2019) ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res 47:D711–D715
69. Edgar R, Domrachev M, and Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210
70. Schmidt EE, Pelz O, Buhlmann S, et al (2013) GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. Nucleic Acids Res 41:D1021-6
71. Tryka KA, Hao L, Sturcke A, et al (2014) NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res 42:D975-9
72. Quezada H, Guzmán-Ortiz AL, Díaz-Sánchez H, et al (2017) Omics-based biomarkers: current status and potential use in the clinic. Bol Med Hosp Infant Mex 74:219–226
73. Olivier M, Asmis R, Hawkins GA, et al (2019) The Need for Multi-Omics Biomarker Signatures in Precision Medicine. Int J Mol Sci 20
74. Serra A, Galdi P, and Tagliaferri R (2019) Multiview learning in biomedical applications, In: Artificial intelligence in the age of neural networks and brain computing, pp. 265–280 Elsevier
75. Fan Z, Zhou Y, and Ressom HW (2020) MOTA: Network-Based Multi-Omic Data Integration for Biomarker Discovery. Metabolites 10
76. Nicora G, Vitali F, Dagliati A, et al (2020) Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Front Oncol 10:1030
77. Lin E and Lane H-Y (2017) Machine learning and systems genomics approaches for multi-omics data. Biomark Res 5:2
78. Serra A, Fratello M, Fortino V, et al (2015) MVDA: a multi-view genomic data integration methodology. BMC Bioinformatics 16:261
79. Pavlidis P, Weston J, Cai J, et al (2001) Gene functional classification from heterogeneous data, In: Proceedings of the fifth annual international conference on Computational biology - RECOMB '01, pp. 249–255 ACM Press, New York, New York, USA
80. El-Manzalawy Y, Hsieh T-Y, Shivakumar M, et al (2018) Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. BMC Med Genomics 11:71
81. EL-Manzalawy Y (2018) CCA based multi-view feature selection for multi-omics data integration. BioRxiv
82. Wang Z, Yuan W, and Montana G (2015) Sparse multi-view matrix factorization: a multivariate approach to multiple tissue comparisons. Bioinformatics 31:3163–3171
83. Ohno-Machado L, Sansone S-A, Alter G, et al (2017) Finding useful data across multiple biomedical data repositories using DataMed. Nat Genet 49:816–819
84. Perez-Riverol Y, Bai M, Veiga Leprevost F da, et al (2017) Discovering and linking

public omics data sets using the Omics Discovery Index. Nat Biotechnol 35:406–409
85. Sun X, Pittard WS, Xu T, et al (2017) Omicseq: a web-based search engine for exploring omics datasets. Nucleic Acids Res 45:W445–W452
86. Khomtchouk B, Vand KA, Wahlestedt T, et al (2016) PubData: search engine for bioinformatics databases worldwide. BioRxiv

Figure Captions

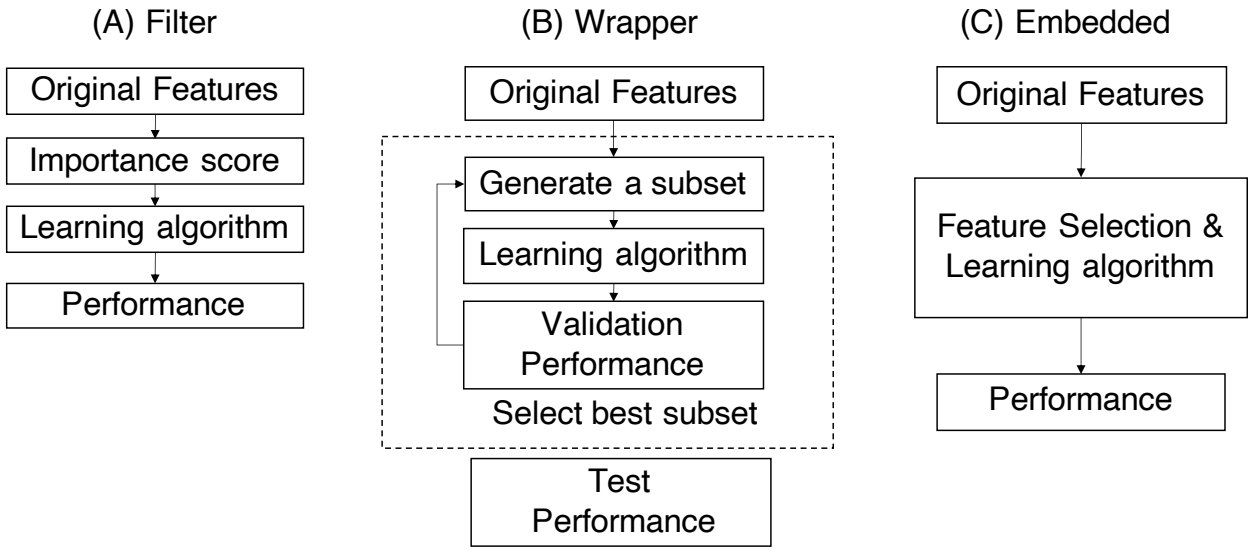

Figure 1 - Filter, wrapper and embedded strategies for feature selection.

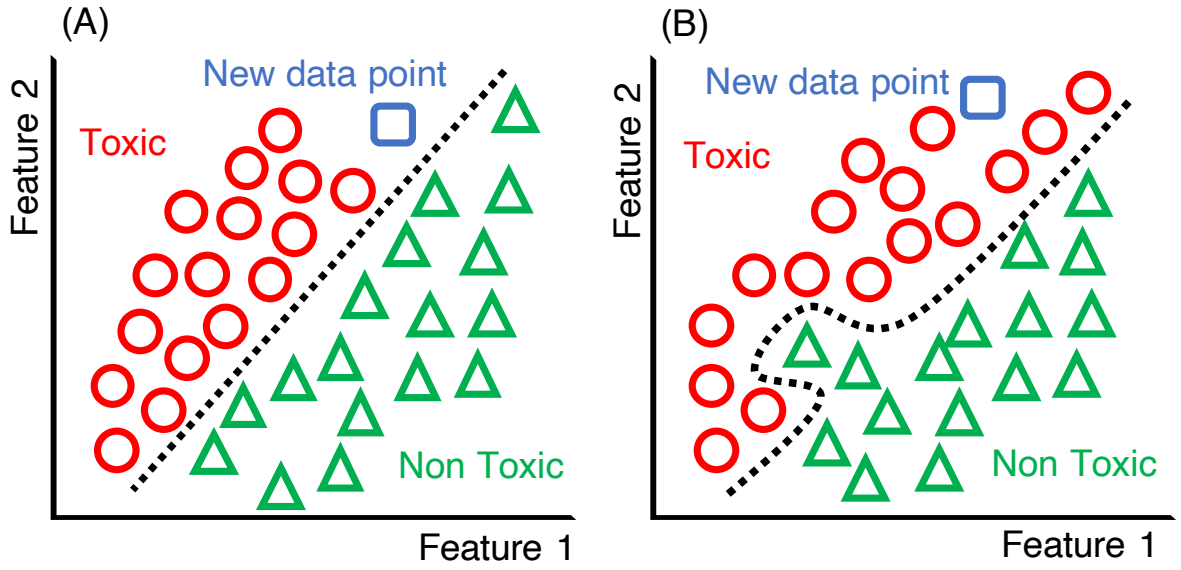

Figure 2 - A simple example of binary classification with toxic (circle) and non toxic compounds (triangle). (A) The model is able to identify a simple relationship between the features of each compound and the classes and to identify a linear boundary between the two classes (dashed line). (B) The model is able to identify a more complex relationship between the features of each compound and the classes and to identify a non-linear boundary between the two classes (dashed

line). In both cases, when a new compound needs to be classified (square), the model will use its feature to estimate to which class it belongs.

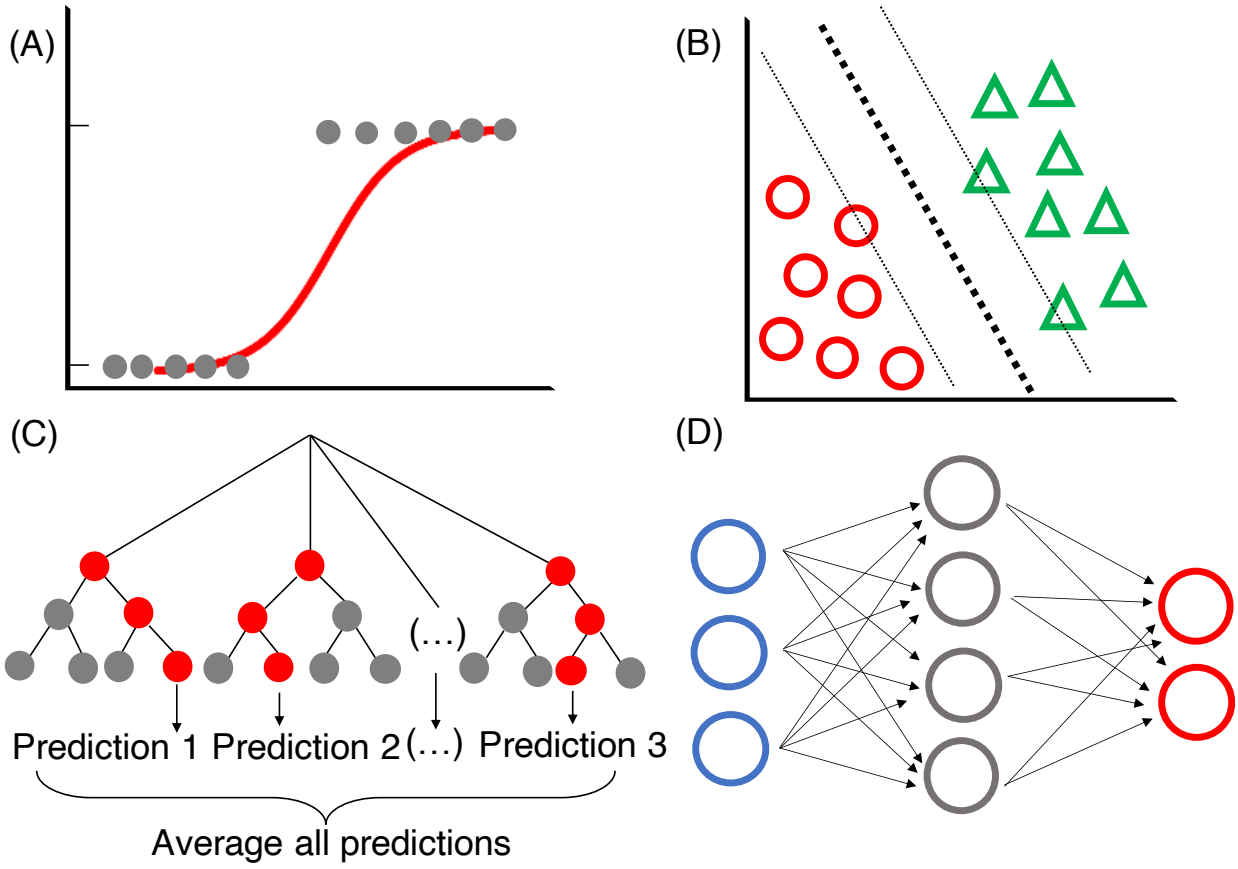

Figure 3 - (A) Example of the logistic regression model. Logistic regression is a well-known method to fit models for categorical data especially for binary responses since it can directly predict probability values restricted in the interval *[0,1]*. (B) Example of linear SVM. The dashed thick line represents the hyperplane that separates the two classes. The space between the two thin lines, called margin, represents the distance between the two classes. Data points following on the margin lines are called support-vectors and are those points that have more impact on the position of the hyperplane. (C) Example of Random forest (RF) classifier. Red dots show the decision paths for a particular data point in each decision tree. From each tree, a prediction is made and the final prediction is computed as the average of all the predictions (D) Example of ANN with three units

in the input layers (blue circles), four units in the hidden layer (grey circles) and two units in the output layer (red circles). The ANN is fully connected since each unit is connected to all the others in the next layer.

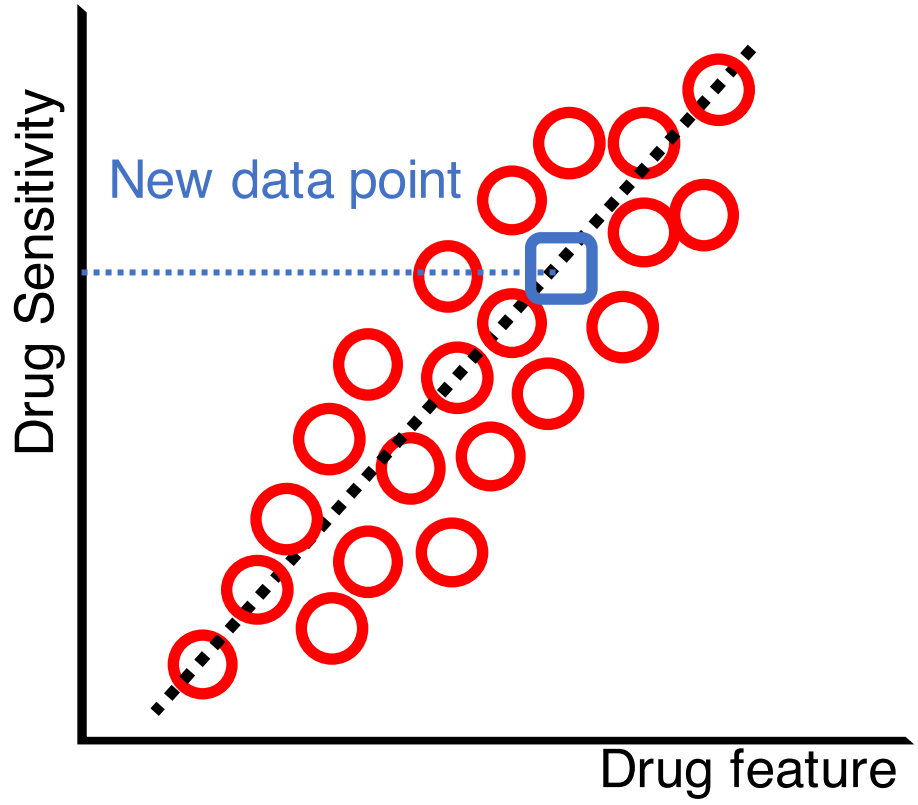

Figure 4 - An example of linear regression to model drug sensitivity. Red circles are drugs represented by their feature values and their sensitivity score. The linear regression model learns a function (dashed line) between the feature values and the sensitivity score. The function is then used to predict the sensitivity score of a new compound.

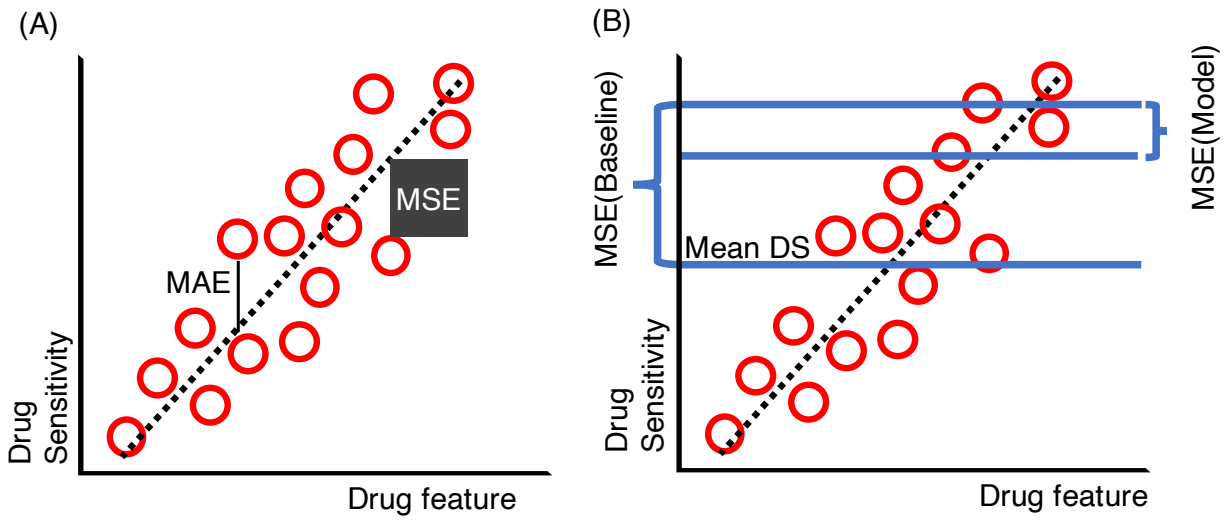

Figure 5 - evaluation metrics for regression. (A) *MAE* measures the absolute distance between the real drug sensitivity value and the predicted one, while the *MSE* measures the square of distance between the true and predictive value. (B) R^2 measure the *MSE* between the real and predicted drug sensitivity value divided the *MSE* of the real and the mean drug sensitivity value.

Table

Table 1 - Example of a confusion matrix for a binary classification problem of toxic *vs.* non toxic chemicals. Rows specify real classes while columns specify predicted classes (~).

	~Toxic	~Non Toxic
Toxic	*25* (*TP*)	*5* (*FN*)
Non Toxic	*10* (*FP*)	*60* (*TN*)

Table 2 - List of public search engines that can be used to link similar omics datasets.

Search Engine	**Description**	**Reference**

Datamed www.datamed.org	It discovers data sets across repositories or data aggregators. It collects different data types including omic, imaging and clinical data.	***(83)***
OmicsDI www.omicsdi.org	A knowledge discovery framework across heterogeneous data (genomics, proteomics, transcriptomics and metabolomics).	***(84)***
Omicseq www.omicseq.org	A web-based platform that facilitates the easy interrogation of omics datasets holistically to improve 'findability' of relevant data.	***(85)***
PubData www.pubdata.bio	It uses novel natural language processing and artificial intelligence algorithms to discover omics datasets worldwide.	***(86)***