Discovery of pairwise biomarkers for dengue severity

Iryna NIKOLAYEVA, Urszula CZERWINSKA, Kevin BLEAKLEY, Anavaj SAKUNTABHAI, Benno SCHWIKOWSKI
Systems Biology Laboratory, Institut Pasteur and Functional Genetics of Infectious Diseases Unit, Institut Pasteur, Paris

Objective

We aim to identify prognostic **biomarkers** from clinical **omics data**. They would **predict** upon arrival at the hospital whether a dengue patient will develop severe dengue.

Is it possible to **combine measurements to make more sensitive and specific biomarkers, compared to single-variable biomarkers?**

Background

Biomedical motivation

Dengue is an emerging tropical viral disease. Infected patients can have very different reactions to the virus. Some have no symptoms, others have a potentially deadly form, dengue shock syndrome (**DSS**).

During virus outbreaks, hospitals can be overcrowded with patients; those who will develop DSS need to be treated very fast. **If we can predict which patients will develop DSS, we can focus resources on them.** We thus search for measurable indicators of DSS upon arrival at hospital. These indicators are also called **biomarkers.**

The dataset

PBMC **blood cells** of 48 Cambodian patients

13 DSS patients

35 non-DSS patients

Measurements for each patient:
67528 **mRNA** expression array values
13 **biochemical** parameters (example: lipids)

Prior results

Earlier analyses showed that the lipids are the best known predictors of DSS. They are easily accessible but not specific and sensitive enough for clinical use.

Approach

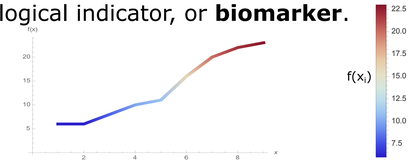
We search for a monotonic regression function of two variables that has the best predictive performance. This regression generalises linear and logistic regressions, while keeping the model constrained to avoid overfitting.

Monotonic functions of one variable

Biologists often manipulate qualitative knowledge: "the higher/lower an indicator, the worse the phenotype". This is equivalent to find f and i such that

$$\varphi = f(x_i)$$

φ is the phenotype
f is a **monotonic function** and
x_i is the biological indicator, or **biomarker**.



Monotonic functions of two variables

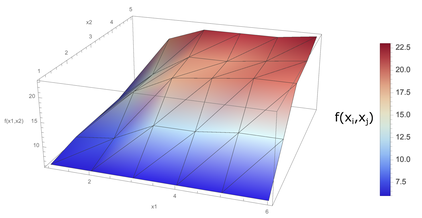
In our case, one variable is not good enough to predict φ (results not shown). We thus slightly complexify the model: We extend the biomarker to two dimensions.
We search for a monotonic function f of two variables i and j such that:

$$\varphi = f(x_i, x_j)$$

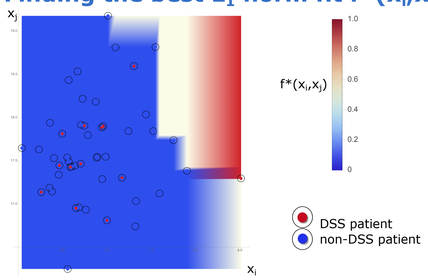
A real-valued function $f(x), x \in \mathbb{R}$, is called *monotonic* if, for any given $i \in 1, \ldots, n$ and any $\Delta \in \mathbb{R}$, the sign of

$$f(\ldots, x_i, \ldots) - f(\ldots, x_i + \Delta, \ldots)$$

does not take on both -1 and 1 over the domain $(x_1, \ldots, x_n) \in \mathbb{R}^n$.



Finding the best L_1 norm fit $f^*(x_i, x_j)$

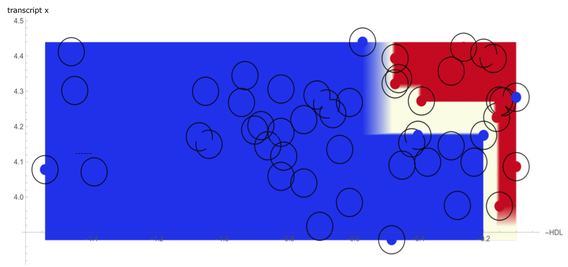


For each pair of variables (i,j), we find the best monotonic fit using the L_1 measure to estimate regression error.

We **assess the performance** of each pair by doing a leave-one-out **cross validation**; we take out one patient at a time and evaluate the probability that it is misclassified.

Results

Best combination of an existing predictor with an additional variable

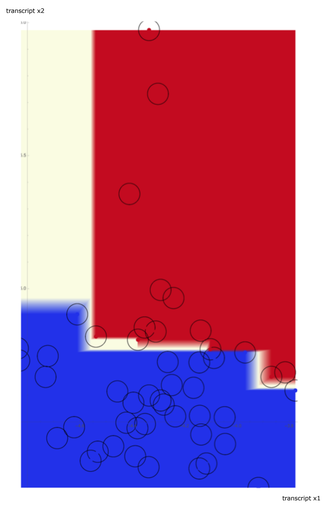


DSS patient
non-DSS patient

Predicted DSS zone
Predicted non-DSS zone

Cross-validation misclassification probability: 0.1

Best pair of variables



Cross-validation misclassification probability: 0.04

Discussion

We have found pairs of biomarkers that are able to predict which patients will develop severe dengue in our data. Nevertheless important concerns remain:

- Are we **overfitting?**
- Can we **replicate** our findings in an other dataset?

Next steps

- Create a p-value that will take into account the predictive performance of each variable.
- Replicate findings on other datasets, that are available in the literature and in our laboratory.

Acknowledgements
Institut Pasteur, LabEx IBEID,
Doctoral school Frontières Du Vivant
Contact iryna.nikolayeva@pasteur.fr