UNIVERSITÉ DE GENÈVE

Département d'informatique

FACULTÉ DES SCIENCES
Professeur R. Appel

Département de Science des
Protéines Humaines

FACULTÉ DE MÉDECINE
Professeur J.-C. Sanchez

Panels of Biomarkers to Improve Patient Classification in Brain Diseases

THÈSE

Présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention interdisciplinaire

par
Xavier Robin

de
Lancy (GE)

Thèse N° 4472

GENÈVE
2012

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

Doctorat ès sciences
Mention interdisciplinaire

Thèse de *Monsieur Xavier Arnaud ROBIN*

intitulée :

" Panels of Biomakers to Improve Patient Classification in Brain Diseases "

La Faculté des sciences, sur le préavis de Messieurs J.-C. SANCHEZ, professeur associé et directeur de thèse (Faculté de médecine, Département de biologie structurale et bioinformatique), R. D. APPEL, professeur ordinaire (Département d'informatique), D. HOCHSTRASSER, professeur ordinaire (Section des sciences pharmaceutiques et Faculté de médecine, Département de biologie structurale et bioinformatique), M. MÜLLER, docteur (Département d'informatique) et J. COLINGE, docteur (Research Center for Molecular Medicine, Austrian Academy of Sciences, Vienna, Austria), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 10 octobre 2012

Thèse - 4472 -

Le Doyen, Jean-Marc TRISCONE

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

À ma famille

À mes amis

Abstract

Biomarkers are characteristics that can be measured objectively in a sample and allow the classification of this sample in two or more groups. In medical applications, they are typically employed to diagnose a disease (or more often to help the physician to establish the diagnosis together with other diagnostic tools), to monitor a treatment, or to make a prognosis about the future outcome of the patient.

Prof. Sanchez' group is focused on protein biomarkers, especially in the field of brain injuries. Several candidates were discovered with proteomics techniques, and later validated with the enrollment of clinical cohorts. Many of these proteins have the potential to be translated to clinical practice. However, their discrimination power (measured with sensitivity and specificity) is not perfect, and research is still ongoing to find better biomarkers.

This research is mostly characterized by technological advances that are continuously emerging in biomedical research. Another approach that has been slowly emerging is to make use of the computational power available nowadays to combine or model the information of several biomarker into a so-called *panel* with better discrimination characteristics.

This thesis focuses on this second approach and presents an original method to the combination of biomarkers. To maximize its acceptance by the medical community we investigated transparent methods that create easily interpretable models. We especially focused on threshold-based combinations, that are the closest to the traditional interpretation of a single biomarker. Other standard methods were also investigated.

Such a project involves thorough statistical analysis. We evaluated both biomarkers and panels with three related measures: receiver-operating characteristic (ROC) analysis, specifically the full and partial areas under the ROC curve (AUC); and sensitivity and specificity measures. To perform this analysis, we developed the pROC package for the R and S+ statistical environments.

An important part in the discovery of a biomarker is the validation step, to ensure that the finding is not a false positive. This step becomes even more crucial with panels of biomarkers, as the risk to over-fit the data grows with the dimension of the search space. To that end we implemented a 10-fold cross-validation scheme allowing the comparison of different combination methods, as well as with single biomarkers.

In order to be usable by scientists in the lab and other researchers without a background in computer science, we implemented this pipeline in a web interface called PanelomiX. It will ultimately help spreading panel analysis of biomarkers and ensure that a robust statistical analysis is performed.

Résumé en français

Les bio-marqueurs sont des caractéristiques qui peuvent être mesurées objectivement dans un échantillon, et permettent, dans le cas d'une classification binaire ou multi-classes, l'assignation de cet échantillon à l'un des groupe. Dans le domaine médical, les bio-marqueurs sont utilisés pour poser le diagnostic d'une maladie (ou, le plus souvent, pour aider le médecin à établir son diagnostic en combinaison avec d'autres outils), pour réaliser un suivi thérapeutique, ou pour faire un pronostiquer les conséquences à long terme pour le patient.

Le groupe de recherche du Prof. Sanchez s'intéresse aux bio-marqueurs protéiques, en particulier dans le domaine des maladies cérébrales. Plusieurs bio-marqueurs candidats ont été découverts au moyen des techniques de protéomiques, puis validés sur de grandes cohortes de patients. Beaucoup de ces protéines ont le potentiel pour être transférées dans la pratique médicale. Cependant, leur puissance de discrimination (mesurée par des statistiques telles que la sensibilité et la spécificité) n'est pas parfaite, et les recherches continuent pour trouver de meilleurs bio-marqueurs.

Cette recherche est caractérisée par l'émergence presque continue de nouvelles innovations technologiques. Une approche alternative est d'utiliser la puissance de calcul disponible grâce aux ordinateurs actuels, afin de combiner (ou modéliser) l'information contenue dans plusieurs bio-marqueurs en un *panel* plus discriminant.

Cette thèse étudie cette deuxième approche et présente une méthode originale de combinaison des bio-marqueurs. Afin d'en maximiser l'acceptation par la communauté médicale, nous avons particulièrement mis l'accent sur la transparence des méthodes de combinaison et la facilité d'interprétation des modèles ainsi créés. En particulier, nous nous sommes intéressés à la combinaison par seuils, qui s'approche au plus près de l'analyse d'un bio-marqueur seul. D'autres méthodes classiques ont également été investiguées.

Un tel projet comprend une part importante d'analyse statistique. Nous avons évalué les bio-marqueurs et les panels à l'aide de trois mesures reliées : la courbe d'efficacité du récepteur (courbe ROC), et en particulier les aires complètes ou partielles sous

cette courbe (AUC), ainsi que les mesures de sensibilité et spécificité. Afin de réaliser cette analyse, nous avons développée pROC, un *package* pour les environnements statistiques R et S+.

Une étape importante dans la découverte d'un bio-marqueurs est sa validation, qui permet de s'assurer qu'il ne s'agit pas d'un faux positif. Cette étape devient encore plus importante avec un panel de bio-marqueurs, car le risque de surapprentissage augmente avec le nombre de dimensions de l'espace de recherche. Pour éviter cela nous avons implémenté une validation croisée en 10 fois, qui permet une comparaison objective des différentes méthodes de combinaison et des bio-marqueurs seuls.

Afin de pouvoir être utilisé par les scientifiques du laboratoire ou d'autres chercheurs sans bagage informatique particulier, nous avons réalisé une interface Web, nommée PanelomiX. Elle permettra à terme de développer l'analyse combinée des biomarqueurs et de s'assurer qu'une analyse statistique robuste est effectuée.

Acknowledgements

I would like to thank all those who made this thesis successful. With no particular order:

- My family who bore with me for all those years.
- My colleagues of the biomedical proteomics research group. Especially Natacha Turck, who is in charge of the clinical research of the lab, and assisted this project at all stages; Natalia Tiberti for her constant curiosity and constant cheerfulness; Alex Hainard, who started with the trypanosomiasis project; and Nadia Walter and Catherine Fouda, who measured many samples.
- All the collaborations I could make during this thesis, and especially Laszlo Vutskits and Marianne Gex-Fabry, who had a significant part in the analysis of the aneurysmal subarachnoid data (chapter 5); Ioannis Xenarios and Nicolas Guex from Vital-IT, who made the most expensive computations possible; Bastien Chopard and Jean-Luc Falcone, who had a seminal role in the optimization of the code.
- The jury who accepted to read this thesis, Ron Appel, Jacques Colinge and Denis Hochstrasser.
- My thesis supervisors, Jean-Charles Sanchez and his never-ending energy, and Markus Müller, never short of an idea whenever I am.
- The Internet, that has the answer to (nearly) every question (especially the trivial, stupid ones).
- And finally thank you, the reader. As Schrödinger outlined (at a small scale approximation) something does only half exist (or rather, has only half the chance to exist) before it is observed. Your reading is therefore absolutely vital!

Organization of this thesis

Chapter 1 is a general introduction on the combination of biomarkers. It offers an outline of the statistical measures applicable to the analysis of biomarkers, an overview of the methods available to combine them into panels, along with relevant definitions.

The review of chapter 2 was published in *Expert Review of Proteomics* in 2009. It presents the combination methods in more detail, with more examples coming from the literature. It analyses the weaknesses of the analysis commonly performed with biomarkers, and proposes a few potential target to improve the robustness of the statistical evaluation of biomarkers.

Chapter 3, published in *BMC Bioinformatics* in 2011, describes the first tool developed to achieve the goals outlined in chapter 2. pROC is a package of tools for the R and S+ statistical environments to perform ROC analysis. It proposes an improved workflow with confidence intervals computation and statistical comparison between two ROC curves, enabling the proper comparison of the performance of biomarkers.

In chapter 4, the main results of this thesis, PanelomiX, both a workflow and a tool, is described in detail. The workflow is organized around the combination of biomarkers based on thresholds found with exhaustive search and optionally pre-filtered with Random Forest. Cross-validation serves both as a validation of the stability of the panel over small changes in the dataset, and to evaluate the performance of the biomarkers on datasets independent from the training of the panel. Finally, ROC analysis with pROC compares the panels with the separate biomarkers.

Chapter 5 and 6 present two clinical applications of the PanelomiX method on the prediction of outcome after aneurysmal subarachnoid hemorrhage (published in 2010 in *Intensive Care Medicine*) and on the staging of patients with human African trypanosomiasis (published in 2009 in *PLoS Neglected Tropical Diseases*).

Finally, chapter 7 summarizes and discusses the results and proposes some perspectives to further improve the biomarker combination methods.

Table of Contents

1

Introduction

1 A brief overview of biomarkers

1.1 What is a biomarker?

According to the *Biomarkers Definitions Working Group*, a biomarker (also named biological marker) is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention[1]."

The *World Health Organization* gives a slightly more restrictive definition in its clinical application. "A biomarker is any substance, structure or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease[2]."

The Medical Subject Headings (MeSH) database (http://www.ncbi.nlm.nih.gov/mesh) also proposes its definition of a biomarker as "Measurable and quantifiable biological parameters (e.g., specific enzyme concentration, specific hormone concentration, specific gene phenotype distribution in a population, presence of biological substances) which serve as indices for health- and physiology-related assessments, such as disease risk, psychiatric disorders, environmental exposure and its effects, disease diagnosis, metabolic processes, substance abuse, pregnancy, cell line development, epidemiologic studies, etc."

All these definitions correspond to two concepts: first the objective measurement of a parameter in a biological sample, and secondly its application to classify patients[1].

Several types of objective biomarker measurements can be performed on patients. One can measure physiological parameters such as the blood pressure[3], or electrocardiogram[4]. Another more complex type of biomarkers are obtained by imaging techniques such as CT scan[4] or MRI[5]. Questionnaire-based clinical scores where a nurse interviews the patient, such as the World Federation of Neurosurgical Societies score (WFNS)[6] or Glasgow Outcome Scale (GOS)[7] can also be considered as biomarkers, as they rely on the evaluation of the patient with precise definitions, leading to robust and objective assessments. Finally, the kind of biomarker that is the most studied in "omics" research groups is the measurement of the

concentration of proteins or other compounds (such as metabolites or gene transcripts) in biological fluids (blood, urine) or tissue samples collected from patients[4]. This list is far from exhaustive and countless measurements have been proposed as biomarkers.

The target of the measurement of a biomarker is to make a useful prediction about the classification of the patient. Typically, one aims at determining the disease status of the patient (stroke or control[8], early or late stage of a disease[9], etc.), monitoring the efficacy of a treatment[4] or predicting the future outcome of a patient[10]. In any case, using the biomarker values one would like to split the patients in two classes according to the status of interest. For the purpose of the analysis, patients are typically labeled with *class labels* according to a known test that was assessed with certainty, called *gold standard*. Future patients will be classified without the need to measure the gold standard. The rare cases where more than two classes exist is called multiclass classification, and is not discussed in this manuscript.

Since proteomics appeared in the 1990s[11], one of its main field of investigation has been the search for biomarkers[12]. Shotgun proteomics[13,14] is still the most used workflow. In short, the proteins in the sample are cleaved after separation by gel electrophoresis or before separation by liquid chromatography, and then analyzed with mass-spectrometry. Bioinformatics software then identifies the peptides and proteins from the mass spectra[15,16].

This methodology has been applied to many biomarker studies. Listing them all would be out of the scope of this introduction. Regarding brain diseases, cerebrospinal fluid (CSF) has been a biological sample of choice. Zhang *et al.* discovered potential biomarkers of ageing[17]. Relevant to stroke diseases, Lescuyer *et al.* and Burgess *et al.* compared ante- and post- mortem CSF to discover biomarkers of brain injury[18,19]. Plasma, although easier to collect from the patient, is a much more complex sample[20]. Progress has been made to discover biomarkers in this sample [21,22] with either sample fractionation[23], affinity-enrichment of peptides[24] or abundant protein depletion[25].

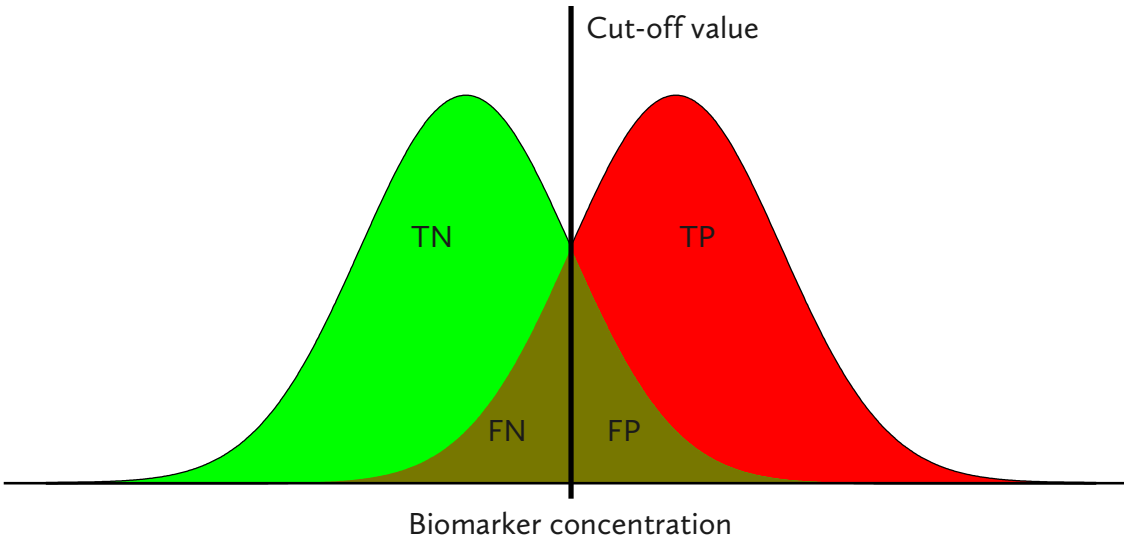


Figure 1: Distribution of control (green) and case (red) patients.

Another popular way to reliably measure the concentration of proteins are immunoassays. Enzyme-linked immunosorbent assays (ELISA) have long been the standard method in this regard[26]. It has two main limitations. First, it requires working antibodies, which are difficult and expensive to produce, limiting their usefulness for screening purposes; second, only one sample can be measured at a time, and as a result more time is required and more sample is consumed when several proteins must be measured. Several methods have been developed to circumvent one or both issues.

1.2 Multiplex technologies for the measurement of protein biomarkers

To cope with the many biomarkers discovered with proteomics techniques, many instruments have been developed to titer several proteins in a biological sample at once using mass spectrometry (MS) or antibody-based techniques.

Mass spectrometry especially when used in shotgun proteomics does not allow the consistent identification (and hence quantification) of a given protein across several samples. In samples where the protein is in lower concentration, the protein may not be identified at all. In addition, mass spectrometry is not a quantitative technique and it is not straightforward to obtain comparable quantitative

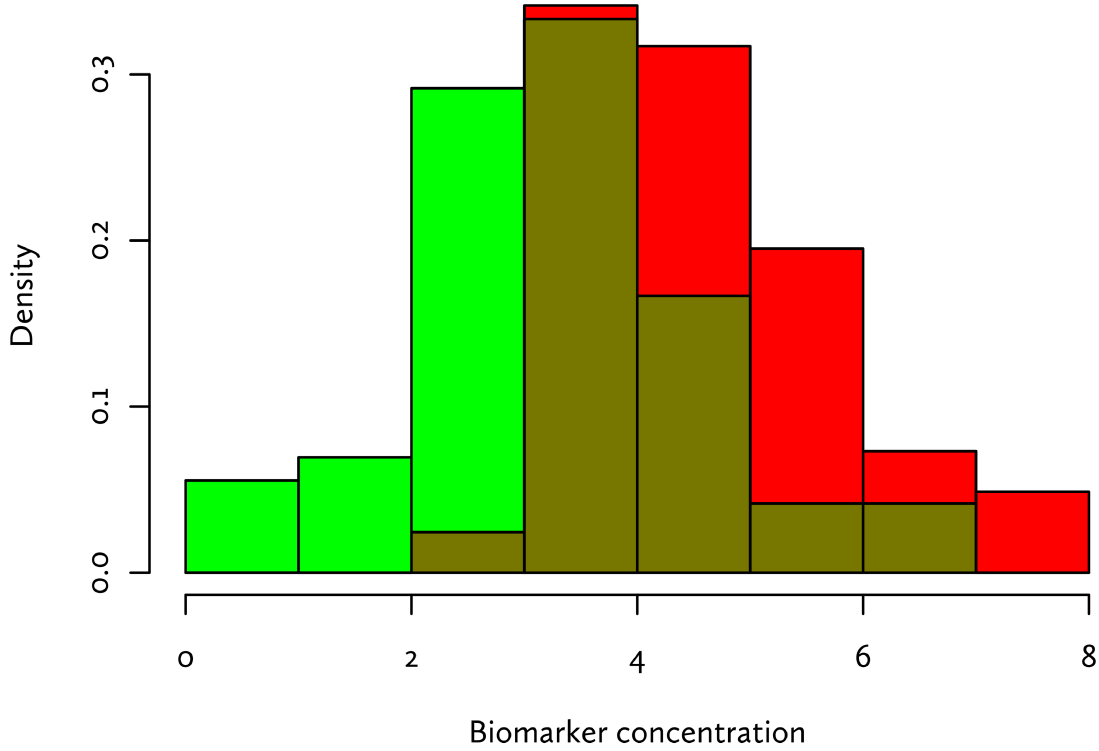


Figure 2: superimposed histograms of the control (green) and case (red) patients.

information from two different samples even in the same conditions[27]. Nevertheless, several techniques have been developed to overcome these issues.

With isobaric labeling with tandem mass tags (TMT) or isobaric tag for relative and absolute quantitation (iTRAQ), peptides are tagged and the samples merged to provide information about the relative concentration between two patients as a ratio. It has been shown that it is also possible to use the reporter-ions as calibration curves to measure an absolute concentration[28]. Label-free mass spectrometry also allows absolute quantification of proteins based on the peak area or height or the spectral counting of identified proteins[29]. Finally, selected reaction monitoring

		Gold standard	
		Positive	Negative
Test outcome	Positive	True Positive (TP)	False Positive (FP)
	Negative	False negative (FN)	True Negative (TN)

Table 1: Contingency table

(SRM), sometimes termed multiple reaction monitoring (MRM), is a targeted MS technique that seeks very specifically for a few peptides in the sample, increasing sensitivity over shotgun techniques, and may become the method of choice for the measure of biomarkers in the next few years[30].

Many immunoassay techniques have been developed to accommodate the measure of several proteins simultaneously with higher throughput. For instance the Biosite Triage Stroke Panel simultaneously measures several biomakers in the blood of patients with ELISA[31,32,33,34]. The Luminex (Luminex Corporation) is a device embedding a flow-cytometer to measure the signal on up to 500 differentially-dyed beads. Antibodies for the different proteins are coated on the beads and the signal can be mapped back to each protein. Other devices such as FlowCytomix (Bender MedSystems)[35] or Bio-Plex (Bio-Rad Laboratories) are based on the same technology.

With planar array assays, such as MULTI-ARRAY (Meso Scale Discovery), A^2 (Beckman Coulter) or FAST Quant (Whatman Schleicher & Schuell BioScience), the antibodies are coated on specific positions of a 2-dimensional array[35].

In short, the measurement of multiple proteins in a single assay is now a routine procedure, even though new developments are still likely to arise. The challenge now resides in the statistical analysis of this amount of data which is the subject of the next section of this manuscript.

		Gold standard		
		Positive	Negative	
Test outcome	Positive	True Positive (TP)	False Positive (FP)	Positive predictive value (PPV) = TP / (TP + FP)
	Negative	False negative (FN)	True Negative (TN)	Negative predictive value (NPV) = TN / (FN + TN)
		Sensitivity = TP / (TP + FN)	Specificity = TN / (FP + TN)	Accuracy = (TN + TP) / (total)

Table 2: Performance measures computation

2 *Statistical evaluation of biomarkers*

To assess its usefulness, a biomarker must be evaluated with statistical methods. The attribution of a patient to the positive or negative class is related to a cut-off (or threshold) as represented in figure 1. With this cut-off, a contingency table can be built (table 1), and performance such as sensitivity and specificity (table 2), which evaluate the usefulness of the biomarker for clinical application, can be measured. When the cut-off is not known *a priori*, a ROC curve can be built to choose the best threshold for a given application or to compare the performance of different markers. Finally, statistical tests will ensure that an apparent good performance is not due to random variations caused by the sampling of the cohort. These tests can be performed either on the ROC curve or on the contingency table.

2.1 Contingency tables

A useful biomarker is a measure whose distribution is different between two classes of patients, with the class labels attributed with a gold standard. This is represented in figure 1. We can see the distribution of biomarker values for the positive patients (red) and the negative patients (green). A cut-off value is represented as the vertical bar. It is chosen to minimize the number of miss-classifications (false positive and false negative patients), represented in brown.

Setting this cut-off allows to define four groups of patients. The true negatives (TN) are the negative patients correctly classified as negatives. They are represented in green on this figure. The true positives (TP), corresponding to positive patients correctly classified as positives, correspond to the red area of the distribution. Both are correctly classified by the biomarker, taken at the given cut-off. Some patients, however, are not correctly classified. They are represented in brown and fall into two classes: the false negatives (FN), who are positive patients incorrectly classified as negatives, and the false positives (FP), negative patients incorrectly classified as positives. Depending on the clinical setting, cut-off values will be chosen to minimize the number of false positives or false negatives or both. The true distribution of the biomarker is generally unknown. An empirical representation of the observed data is the histogram, as shown in figure 2.

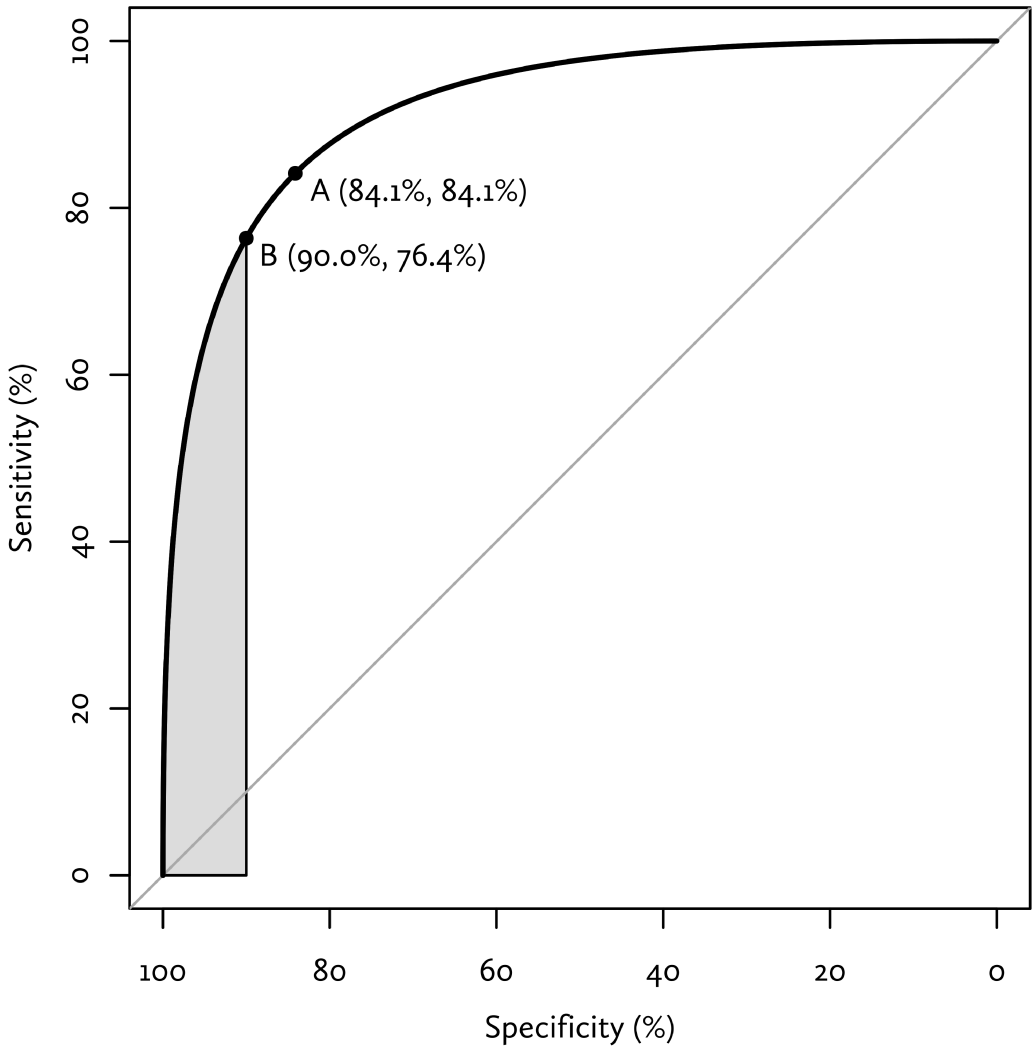


Figure 3: A simple ROC curve corresponding to the data in figure 1. The partial AUC is displayed in grey together with the best sensitivity and specificity achievable in this whole curve (A) or in the partial region of interest (B).

Typically, concentrations of the biomarker below the given cut-off will classify the patient in the negative class, and levels above the threshold mean the patient is in the positive class. However, for some biomarkers this rule can be reversed and lower concentrations are observed in negative patients. In the next sections we assume that higher values correspond to a higher probability for the positive class.

A common way to represent the TP, TN, FP, and FN values is the contingency table as shown in table 1. When filled with the patient counts, it can be used to determine clinically useful performance measures that will be detailed in the next section.

2.2 Sensitivity, specificity and other performance measures

The contingency table shown in the previous section, with the count of true and false positive and negative patients, is not very useful *per se*. However, it is an important step towards the computation of clinically relevant performance measures such as sensitivity (SE) and specificity (SP). Sensitivity is defined as the proportion of positive patients correctly detected by the test. Similarly, specificity is the proportion of negative patients correctly rejected by the test. Both sensitivity and specificity measure how well the test performs to classify a patient. On the other hand, positive and negative predictive values (PPV and NPV) measure the probability for a patient to be actually positive or negative, given the test outcome. They take into account the prevalence of the positive occurrence in the sample, i.e. the total number of positive patients compared of the number of negative ones. Finally, accuracy is the proportion of correctly classified patients within the total sample. All these measures are shown in table 2.

Many other performance measures exist and are computed either from the contingency table, or from the performance measures derived from this table. For example, odds ratio (OR) measures the effect of a given increase of the studied marker. They are computed as (SE/(1-SE))(SP/(1-SP)). The likelihood ratio summarizes how likely positively classified patients are truly positive (positive likelihood ratio, LR+ = SE / (1 - SP)) or negative (negative likelihood ratio, LR- = SP / (1 - SE)), compared with control patients[36].

A totally different approach unrelated with contingency tables is to compute cost-efficiency ratio of the biomarker. For instance, the impact on the health of patients can be measured with scores such as the quality-adjusted life-years[37,38]. This approach requires longer follow-ups of the patients, but gives a more precise indication of the usefulness of the biomarker.

2.3 ROC curves

All the measures presented in the previous section are related to a static cut-off that splits the patients into positive or negative test groups. However, the cut-off value is not always known *a priori*. Even in the case where the optimal threshold value is

known, one may want to tune the trade-off between sensitivity and specificity to adjust to the clinical needs that may require higher values of sensitivity or specificity, at the expense of the other. A receiver operating characteristic (ROC curve) is a plot that displays all possible cut-off values, with the associated sensitivities and specificities. A ROC curve is shown in figure 3, corresponding to the data of figure 1. Point A corresponds to the (sensitivity, specificity) of the cut-off shown in figure 1. It matches the point with the highest Youden's J statistic[39] or the point that lies closest to the perfect classification at 100% sensitivity and specificity[40], with possible adjustments[41].

A common and useful performance measure that can be derived from the ROC curve is the area under the curve (AUC). With empirical ROC curves the AUC can be computed with the trapezoidal rule[42]. It can be shown that a convenient interpretation of the AUC is "the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance[42]." Consequently, higher AUCs correspond to better discrimination of the biomarker between the two classes. An AUC of 0.5 means that the discrimination is random and that the positive and negative distributions completely overlap. Rather than estimating the prevalence and the cost of miss-classifications, a simple way to shift the threshold toward higher sensitivity or specificity is to compute only a part of the AUC (partial or pAUC), limited to the portion of sensitivity or specificity that is acceptable for the application[43].

ROC curves allow to dynamically choose a cut-off based on the clinical trade-offs. For instance, one may shift the cut-off slightly to the left to achieve a better specificity, at the expense of a lower sensitivity. This is the case of threshold B, which has the highest sensitivity for at least 90% specificity. If a high sensitivity is of prime importance, the cut-off can be shifted to the right, at the expense of a lower specificity. This shift can be computed automatically with the prevalance and costs of the different kinds of miss-classifications, for example with the formulas given by Perkins *et al.*[41]. With the partial AUC approach, the threshold will then be chosen to fall within the region of interest.

2.4 Statistical tests

We have seen several measures of biomarker performance. Once they are computed, statistical tests must be performed to ensure that their apparent efficiency is not caused by random variation due to the sampling. Because biological data is generally not normally distributed, non-parametric tests are the method of choice[44]. Two kinds of tests can be performed: the univariate evaluation of the biomarker, and the multivariate comparison of two biomarkers.

Univariate methods look at one biomarker independently. They evaluate if its expression is different between the two studied groups. Mann-Whitney U test[45], a non-parametric equivalent of Student's t-test, is a test to compare the median of a continuous biomarker between the two groups. Indeed, this test is equivalent to the AUC[46]. Finally, Fisher's exact test[47] or similar tests[48] can assess the significance of a contingency table.

Unlike univariate tests, the goal of multivariate tests is to compare the performance of two or more usually correlated (or paired) biomarkers, for example based on the ROC curves. Parametric tests make use of the binormal distribution (two overlapping normal distributions)[49]. While this assumption has been shown to be rather robust to deviations from the normality[50], it seems safer to free ourselves from the binormal distribution. In 1983, Hanley and McNeil proposed a semi-parametric test[51], and in 1988, DeLong *et al.* proposed a fully non-parametric test to compare two or more ROC curves[52]. This latter test is still the method of choice today, and is implemented in a number of software. Other tests have been proposed to compare AUCs[53,54], or the shape of paired[55] or unpaired[56] ROC curves. For partial AUCs, bootstrap or permutation tests can be employed, although they are not widely implemented in statistical software[57].

3 Panels of biomarkers

The basic measurement of the usefulness of a biomarker is its performance estimated with sensitivity and specificity, which must be as high as possible. For instance for Alzheimer's disease, it has been determined that an ideal biomarker should display at least 80% sensitivity and 80% specificity[58]. While this limit of

usefulness is different in the context of each disease, biomarkers analyzed separately often suffer from insufficient sensitivities and specificities. Therefore, panels of biomarkers have been proposed as a potential tool to improve the classification of the patients.

3.1 Basic definition

Combining biomarkers is an application of supervised machine learning. The goal is to build a model integrating multiple inputs (levels of the biomarkers and clinical information) into a single output (the patient classification). The model is learned or trained from known examples, in which the class labels (associating patients to one of the two classes) are known. A useful and efficient model is then able to generalize from the training sample, in order to be able to predict the classes of new patients who have never been seen before[59]. The choice of the method depends on the characteristics of the dataset to be analyzed. Depending on the shape of the class boundaries and on the number of training examples, a method providing a linear separation may or may not be adequate.

Many different combination methods exist. They are described in detail in section 3.3. They differ by several factors: the algorithms applied to transform the input into the output, the method employed to determine the parameters, and as a result the shape of the separation boundaries they are able to achieve. Threshold-based methods associate each predictor included in the panel with a threshold. Each of them corresponds to a separating hyperplane in the marker value space perpendicular to the marker axis where the threshold is applied. Decision trees apply different decisions in a hierarchical tree structure. Regression methods such as linear or logistic regression models apply a regression formula of the form $y = X\beta+\varepsilon$. The parameters β and ε are fitted with least squares or maximum likelihood methods. Finally, more complex methods such as support vector machines or neural networks also exist. They are usually applied to produce more complex separations in space, at the expense of a higher risk of overfitting. Many other methods have been developed over time, and it would be outside the scope of this introduction to describe them

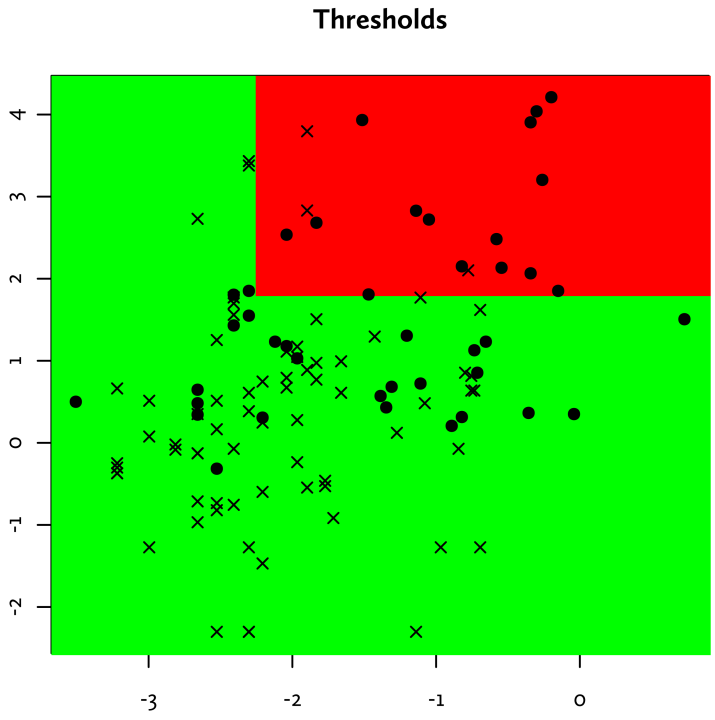


Figure 4: Threshold classification into good (green) and poor (red) outcome after aneurysmal subarachnoid hemorrhage (chapter 5). Crosses and dots represent observed good and poor outcome patients respectively.

all. We will only outline those that have actually been used to combine biomarkers, and show a few examples of their usage for the combination of biomarkers.

Before combining the biomarkers, one needs to select the features or variables that will be included in the model. This step is named feature selection and is discussed in the next section. Sometimes it is performed directly by the model training algorithm and is therefore unattended.

3.2 Feature selection

Feature selection is the selection of the set of biomarkers or *features* that will be included in the panel. It is is of critical importance in high-dimensionality problems such as microarray or mass spectra analysis, but becomes less crucial in lower dimensional biomarker combination studies where only a few markers are studied. Feature selection consists in the selection of a few biomarker or clinical informations to include in the model. The use of feature selection in bioinformatics has been extensively reviewed[60,61].

Feature selection methods can be classified in three categories: filter methods, wrapper methods and embedded methods[60]. Filter methods apply a simple filter on the features and select those that best meet the criterion, disregarding the classification method that will be employed. In wrapper methods, the selection is wrapped directly around the classifier, providing a better accuracy of the feature selection. With embedded methods, the selection is performed directly within the training process. Finally, hybrid methods can also be defined as an intermediary between filter and wrapper methods[62].

Filter selection methods can be further separated in two classes: univariate and multivariate. Univariate filters look at each feature separately, ignoring the correlation and correlations with the other features. As they are unrelated with the classification algorithms, they ignore the dependencies between the features that can have a large effect in the classification, and often produce worse classification accuracy. Multivariate filters on the other hand take into account the interactions

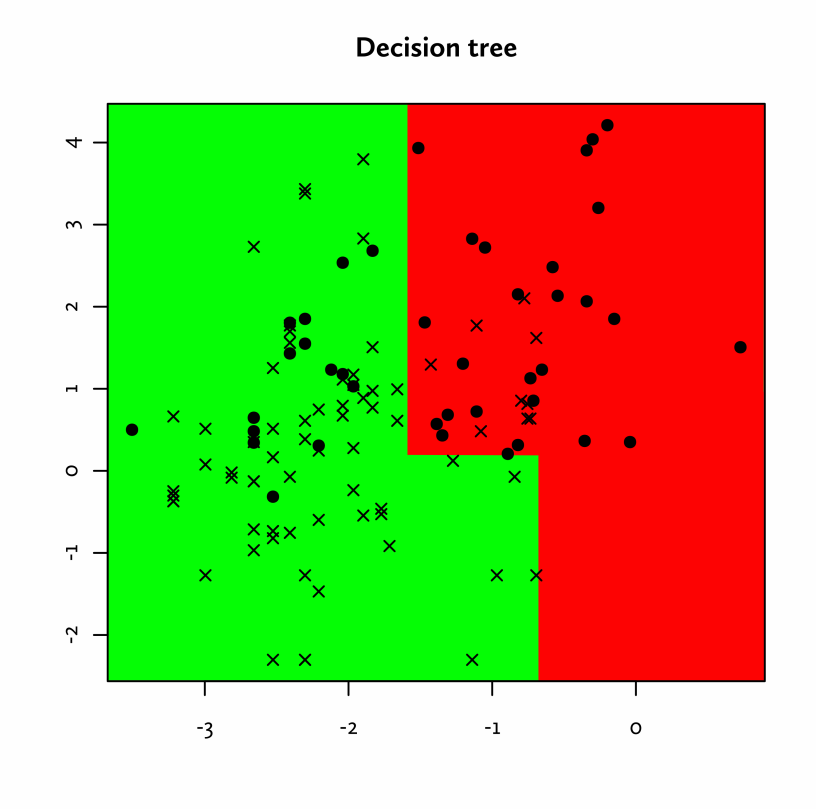


Figure 5: Decision tree classification into good (green) and poor (red) outcome after aneurysmal subarachnoid hemorrhage (chapter 5). Crosses and dots represent observed good and poor outcome patients respectively.

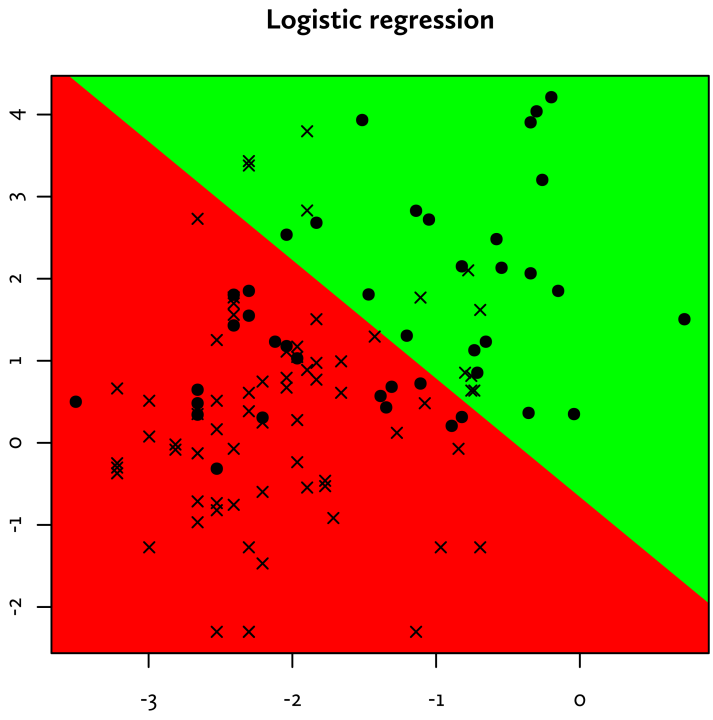


Figure 6: Logistic regression classification into good (green) and poor (red) outcome after aneurysmal subarachnoid hemorrhage (chapter 5). Crosses and dots represent observed good and poor outcome patients respectively.

between the variables, but they still ignore the interactions specific to the classification method[60].

In the field of biomarker combination, Hilario and Kalousis report several applications[61]. Baggerly *et al.* reduced an initial 60 831 m/z peaks from mass-spectrometry to 506 peaks and then five features with various feature selection techniques[63]. Petricoin *et al.* wrapped a genetic algorithm around a self-organizing map for mass-spectrometry[64]. Genetic algorithms are also frequently wrapped around support vector machines (GA/SVM), for instance by Peng *et al.*[65]. Another very popular feature selection technique is random Forest[66,67]. By creating a large number of decision trees, it is possible to extract the most frequent split-points and thus the most interesting variables.

Once the features have been selected, the combination algorithm can be applied. We will now review the main methods available to do this.

3.3 Combination methods

Threshold-based methods

Because the determination of a thresholds (or cut-offs) is the most common way to analyze a biomarker, it is intuitive to apply a similar kind of analysis to panels. With threshold-based methods, a set of thresholds (one for each feature) is selected. The panel value is the count of biomarker measures exceeding the respective thresholds. A more formal definition, together with an algorithm to select the features thresholds, is given in chapter 4. This method is employed mostly with ELISA and clinical data, where the targets are known and measured sufficiently reliably to determine a threshold. Threshold panels have two main advantages: once the threshold values are determined the results are straightforward to calculate, and the simple boundary structure (figure 4) significantly limits the risk to over-fit the data.

The main challenge with this method is to determine the set of cut-offs. This is usually performed in a univariate manner[68,69,70,71], but attempts have been made to select cut-offs with multivariate methods[9,10,72]. For instance, Reynolds *et al.* developed

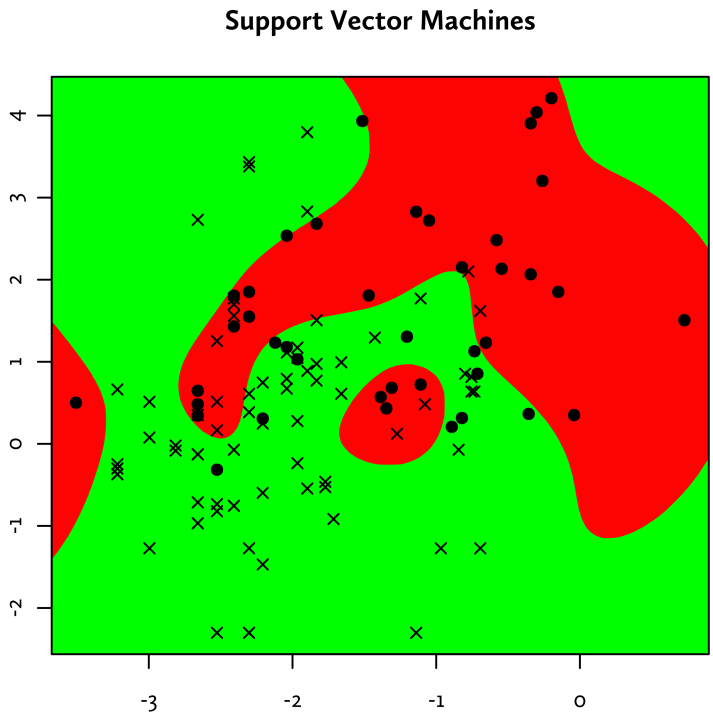


Figure 7: Support vector machines classification into good (green) and poor (red) outcome after aneurysmal subarachnoid hemorrhage (chapter 5). Crosses and dots represent observed good and poor outcome patients respectively.

an iterative algorithm to select the thresholds[72]. A methodology based on exhaustive search with pre-filtering is presented in detail in chapter 4.

Decision trees

Decision trees are quite similar to threshold-based methods in that they determine cut-offs to make binary splits of the biomarkers (figure 5). However, they can create much more elaborate rules to find complex boundaries. More than one threshold can be applied to a biomarker, and not all are necessarily evaluated. This results in the partition of the feature space into boxes, as shown in figure 5. Several tree construction methods exist, which differ in the way the tree is grown from the training dataset (the selection of features and thresholds for each node), and the pruning strategy. The main advantages of decision trees are the ability to directly classify the patients in more than two groups, and to easily combine categorical and continuous variables.

One of the most popular tree method is the classification and regression trees (CART)[73,74,75]. Other methods include C4.5[76,77], J48[78,79] or recursive partitioning (Rpart)[80,81].

Several related methods have been employed for panels. For instance, the patient rule-induction method (PRIM) takes a slightly different approach. Like decision trees it creates simple boxes in the feature space. However, PRIM maximizes the proportion of positive patients in the box rather than the accuracy of the classification[59,82,83,84,85]. Multivariate adaptive regression spline (MARS) models are intermediary between decision trees and linear regressions that defines piecewise linear functions. It was chosen by Warner *et al.* to combine biomarkers of potential harm in cigarette smokers[86], and by Brasier *et al.* to develop a candidate panel for the detection of dengue hemorrhagic fever[87].

Decision trees can be combined with boosting algorithms to improve the classification results. A famous example of this integration is the random forest algorithm[88]. For example, it has been applied to the prediction of dengue hemorrhagic fever[89] and for colorectal cancer[90]. As mentioned previously, they also frequently serve to select features in large datasets[66,67].

Linear regression methods

Logistic regression is especially popular in medical research, where it is widely employed to model risk factors rather than as a combination method for the classification of new patients[91,92]. It is appreciated for its simplicity and robustness. It is based on a clear statistical formulation, and the solution is globally optimal. It can deal with interactions to model nonlinear class boundaries, but this requires a priori knowledge about the structure of the data. The resulting separation is a straight, diagonal line as shown in figure 6.

Logistic regression can combine continuous or categorical data, either biomarkers or clinical information. For example, Visintin *et al.* combined only biomarker data[93], while Welsh *et al.* combined protein biomarkers with clinical data[94] and Wicki *et al.* combined clinical information only[95]. Logistic regression is often employed together with other combination methods. For instance, Reynolds *et al.* and Montaner *et al.* used both logistic regression and a threshold-based method[70,72]. Similarly Reddy *et al.* compared the performance of the Logical Analysis of Data (LAD) methodology with logistic regression and other methods[76].

Several methods are derived from logistic regression. For instance, Nolen *et al.* applied additive logistic regression for the early detection of lung cancer[96], a method inspired by boosting[97]. Another algorithm is the Metropolis-Monte Carlo method, which has been applied by Yurkovetsky *et al.* to the early detection of ovarian cancer[98], and by Nolen *et al.* for lung cancer, among other methods[96]. The score is a linear combination of the biomarkers, with coefficients estimated with Monte Carlo optimization.

Support vector machines

Support vector machines (SVM) are among the most popular machine learning methods, providing a clear mathematical model with a globally optimal solution, contrary to many other methods which can get trapped in local optima. While the class separation is linear, it is applied after a kernel transformation, enabling complex separation boundaries (figure 7). They are mostly employed to classify high dimensional datasets, for instance microarrays[99,100,101,102] or mass spectrometry (MALDI[75,103,104], SELDI[76,78,105]or ESI[106]). For instance Prados *et al.* created classifiers of

SELDI spectra[78] to diagnose patients after a stroke event. More recently Frenzel *et al.* applied SVM to MALDI-ToF spectra to predict the outcome of patients with acute lung injury or acute respiratory distress syndrome[75]. In both cases, no precise biomarker was identified, even though Prados *et al.* could target the few especially interesting peaks[78]. Other efforts have also been made in this domain[107].

SVM are normally not employed to combine a small number of validated biomarkers. An exception to this rule is the work by Wild *et al.* who applied SVM to combine ELISA data in order to classify patients suffering from rheumatoid arthritis[108]. However, they used SVM only to challenge the regularized discriminant analysis.

A strong expertise is required for the successful application of SVM. The choice of the kernel type and parameters as well as error penalties is critical to avoid over-fitting and produce a generalizable model[109]. These parameters depend on the dataset itself and cannot be pre-set in a software. Therefore, a nested cross-validation scheme[110] or other bias-correction methods[111] must be applied.

Neural networks

Artificial neural networks (ANN) are powerful classification models that can combine biomarkers in a nonlinear manner. They are based on the association of network units with weights into a series of input, hidden and output layers with an algorithm called perceptron[59]. While they are very powerful models, it is difficult to extract biological knowledge from them.

Examples of ANN usage include the detection of women with high risk to develop an ovarian cancer by Zhang *et al.* and Donach *et al.*[112,113], the and detection of lung cancer by Nolen *et al.* and Flores-Fernández *et al.*[96,114].

Others

Several other supervised classification methods exist. With naive Bayes classifiers, the probability to belong to a class is computed based on parameters estimated independently on all the predictors[115]. It has been applied by Ralhan *et al.* to detect patients with head-and-neck squamous cell carcinomas with proteins identified with mass spectrometry after labeling with isobaric tandem mass tags[116]. Nolen *et al.*

applied it together with other methods to combine 81 serum proteins for the early detection of lung cancer[96].

Regularized discriminant analysis (RDA) is a linear combination method that can deal with strongly correlated data[59]. Wild *et al.* applied RDA to combine three markers in a panel for detecting patients with rheumatoid arthritis[108]. Flores-Flores-Fernández *et al.* appied LDA in in comparison with a neural network approach[114] and Wu *et al.* with several other classification methods[103].

Together with their Triage Stroke Panel, Biosite developed the Multimarker Index (MMX), a proprietary algorithm to combine the four measured biomarkers into a single result[33,34].

Torkaman *et al.* developed an innovative approach based on cooperative game theory[77]. They employed this new method to classify different types of leukemia.

Cox proportional hazards model is especially well suited to survival analysis. For instance, Ring *et al.* combined biomarkers for the diagnostic of estrogen receptor-positive breast cancer[117], and Damman *et al.* predicted the mortality after heart surgery[118].

Knickerbocker *et al.* combined both protein microarray and clinical data of patients after renal replacement with generalized additive models (GAM)[119]. The same method was also applied by Brasier *et al.* to diagnose dengue hemorrhagic fever[89].

As we can see, a large variety of combination methods are available. The choice of one or another is essentially a matter of personal preference, combined with the knowledge of the data. In the end, it is possible to build a model that displays an exquisite performance on the training data. However, it is only useful if it can be generalized on independent datasets. The validation step makes sure it is the case.

3.4 Validation

The last step in the statistical validation of a panel of biomarkers is to evaluate its performance on an independent dataset, to avoid the over-fitting, namely the over-optimism that follows the evaluation of the performance of a model on the same dataset that was employed to train it (also designed as reclassification). While the

best possible validation is to collect an independent test cohort[62], preferentially in a randomized controlled trial to assess the impact on the health of patients[38], it is not always possible to achieve this due to time and cost constraints[120]. Fortunately, several computational methods make it possible to obtain an unbiased evaluation of the performance of a model even without a dedicated test set[121].

Cross-validation is a purely computational method. It functions by splitting the dataset into k parts of equal size, and sequentially using k-1 parts to train the model while keeping one part aside as test set. The performance estimate is the mean of the performance of the individual k parts. The main drawback of this method is that the size of the training set is slightly reduced, resulting in a worse classification model, overestimating the classification error in some scenarios. In addition, the learning step must be repeated k times, which can prove computationally intensive with high values of k. The most extreme case is leave-one-out cross-validation with $k = n$ (the number of patients) requiring to apply the learning method n times but providing a nearly unbiased estimate of the method's performance[59].

Bootstrap is another purely computational method. Instead of splitting the data, it randomly selects observations with replacements, thus generating a new sample of the same size as the original one. Approximately 63% of the sample is selected, leaving 37% of the observations as test set, independent from the 63% employed to train the model. However, the training sample contains repeated measures and complex correction methods must be applied[59].

If the sample size is sufficient, it is possible to leave a subset of the sample aside and use it as a validation set[72,93]. As for the cross-validation, this means that the size of the training set is significantly reduced. Consequently, the training of the model is not optimal, resulting in pessimistic performance estimates. In addition, fewer observations are available in the test set, leading to a less precise estimate of the performance of the model.

Finally, permutation tests assess whether the results of the classification are significant or not. The labels of the patients are randomly shuffled and the training method is applied. If the classification results on the random dataset are comparable

to those obtained in the original sample, it is a strong indication that the classifier over-fits the data[122,123]. However, this method does not allow estimating the true classification performance, and therefore its usefulness is rather limited.

As a result, cross-validation is the most popular validation method in bioinformatics[59,124], although it is often used in a non-standard manner[124].

4 Goals of the thesis

In this thesis, we hypothesized that the information about the class of the patient was complex and could not be deduced from the measurements of a single biomarker. Therefore, combining several biomarkers into a single, multiplexed output could lead to an improvement in the classification of the patients. Machine learning techniques will be applied to datasets relevant in the lab, especially related to brain diseases.

The specific goals of this thesis were three fold.

4.1 Propose a framework to easily create white-box panels of biomarkers

Early discussions with medical practitioners quickly made it evident that a clear, understandable way to combine biomarkers would be preferred over a mathematically more complex approach. The reasons are two-fold:

1. Understanding the underlying function of a system is the goal of all biologists. Medical practitioners also want to understand how the patients function.
2. Previous research with black-box approaches[64] have turned to be highly ineffective when applied to independent datasets[125]. While the reason may lie in the analytical procedures rather than in the combination method in itself, a transparent approach where every component of the model can be analyzed and understood would make it easier to detect such bias.

Therefore, the main goal of this thesis is to develop an approach where the biomarkers are combined in a transparent, interpretable way.

4.2 Study the performance of the proposed panels

Once the method and the panels have been established, the second task is to ensure that the panels and the methods are efficient enough. We need to ensure that the panel provides an improvement over the biomarkers taken individually, and also that it provides reasonable performance compared to other combination methods (white- or black-box). Comparing panels with its constituting biomarkers is usually not performed in most biomarker combination studies and we will have to find the most accurate way to do it.

4.3 Build interfaces to be used by the scientists in the lab

While programming and command-line interfaces are powerful tools for computer scientists and bioinformaticians, they are worthless to anyone lacking programming skills. My special position of embedded bioinformatician within a wet lab gives me the responsibility to make all the tools available to those researchers who do not have the technical skills to use command-line or programmable bioinformatic tools. Therefore, the most important tools developed during this thesis will be made available with graphical user interfaces (GUI).

5 References

1. Biomarkers Definitions Working Group, Atkinson A. J., Colburn W. A., *et al.*, (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69 (3), p. 89-95.

2. World Health Organization, (2001). *Biomarkers in Risk Assessment: Validity and Validation*, Geneva.

3. Strimbu K. & Tavel J. A., (2010). What are Biomarkers? *Current opinion in HIV and AIDS*, 5 (6), p. 463-466. DOI: 10.1097/COH.0b013e32833ed177.

4. Vasan R. S., (2006). Biomarkers of Cardiovascular Disease Molecular Basis and Practical Considerations. *Circulation*, 113 (19), p. 2335-2362. DOI: 10.1161/CIRCULATIONAHA.104.482570.

5. Dickerson B. C., (2010). Advances in quantitative magnetic resonance imaging-based biomarkers for Alzheimer disease. *Alzheimer's Research & Therapy*, 2 (4), p. 21. DOI: 10.1186/alzrt45.

6. World Federation of Neurological Surgeons Committee, (1988). Report of World Federation of Neurological Surgeons Committee on a Universal Subarachnoid Hemorrhage Grading Scale. *Journal of Neurosurgery*, 68 (6), p. 985-6.

7. Jennett B. & Bond M., (1975). Assessment of outcome after severe brain damage. *Lancet*, 1 (7905), p. 480-484.

8. Hasan N., McColgan P., Bentley P., *et al.*, (2012). Towards the identification of blood biomarkers for acute stroke in humans: a comprehensive systematic review. *British Journal of Clinical Pharmacology*. DOI: 10.1111/j.1365-2125.2012.04212.x.

9. Hainard A., Tiberti N., Robin X., *et al.*, (2009). A Combined CXCL10, CXCL8 and H-FABP Panel for the Staging of Human African Trypanosomiasis Patients. *PLoS Neglected Tropical Diseases*, 3 (6), p. e459. DOI: 10.1371/journal.pntd.0000459.

10. Turck N., Vutskits L., Sanchez-Pena P., *et al.*, (2010). A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Medicine*, 36 (1), p. 107-115. DOI: 10.1007/s00134-009-1641-y.

11. Wilkins M. R., Pasquali C., Appel R. D., *et al.*, (1996). From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis. *Nature Biotechnology*, 14 (1), p. 61-65. DOI: 10.1038/nbt0196-61.

12. Srinivas P. R., Verma M., Zhao Y., *et al.*, (2002). Proteomics for Cancer Biomarker Discovery. *Clinical Chemistry*, 48 (8), p. 1160-1169.

13. Washburn M. P., Wolters D. & Yates J. R., (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, 19 (3), p. 242-247. DOI: 10.1038/85686.

14. Wolters D. A., Washburn M. P. & Yates J. R., (2001). An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Analytical Chemistry*, 73 (23), p. 5683-5690. DOI: 10.1021/ac010617e.

15. Eng J. K., McCormack A. L. & Yates III J. R., (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5 (11), p. 976-989. DOI: 10.1016/1044-0305(94)80016-2.

16. Colinge J., Masselot A., Giron M., *et al.*, (2003). OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 3 (8), p. 1454-1463. DOI: 10.1002/pmic.200300485.

17. Zhang J., Goodlett D. R., Peskind E. R., *et al.*, (2005). Quantitative proteomic analysis of age-related changes in human cerebrospinal fluid. *Neurobiology of Aging*, 26 (2), p. 207-227. DOI: 10.1016/j.neurobiolaging.2004.03.012.

18. Lescuyer P., Allard L., Zimmermann-Ivol C. G., *et al.*, (2004). Identification of post-mortem cerebrospinal fluid proteins as potential biomarkers of ischemia and neurodegeneration. *PROTEOMICS*, 4 (8), p. 2234-2241. DOI: 10.1002/pmic.200300822.

19. Burgess J. A., Lescuyer P., Hainard A., *et al.*, (2006). Identification of Brain Cell Death Associated Proteins in Human Post-mortem Cerebrospinal Fluid. *Journal of Proteome Research*, 5 (7), p. 1674-1681. DOI: 10.1021/pr060160v.

20. Anderson N. L. & Anderson N. G., (2002). The Human Plasma Proteome History, Character, and Diagnostic Prospects. *Molecular & Cellular Proteomics*, 1 (11), p. 845-867. DOI: 10.1074/mcp.R200007-MCP200.

21. Jacobs J. M., Adkins J. N., Qian W.-J., *et al.*, (2005). Utilizing Human Blood Plasma for Proteomic Biomarker Discovery†. *Journal of Proteome Research*, 4 (4), p. 1073-1085. DOI: 10.1021/pr0500657.

22. Koomen J. M., Li D., Xiao L., *et al.*, (2005). Direct Tandem Mass Spectrometry Reveals Limitations in Protein Profiling Experiments for Plasma Biomarker Discovery. *Journal of Proteome Research*, 4 (3), p. 972-981. DOI: 10.1021/pr050046x.

23. Faca V., Pitteri S. J., Newcomb L., *et al.*, (2007). Contribution of Protein Fractionation to Depth of Analysis of the Serum and Plasma Proteomes. *Journal of Proteome Research*, 6 (9), p. 3558-3565. DOI: 10.1021/pr070233q.

24. Pernemalm M., Lewensohn R. & Lehtiö J., (2009). Affinity prefractionation for MS-based plasma proteomics. *Proteomics*, 9 (6), p. 1420-1427. DOI: 10.1002/pmic.200800377.

25. Plavina T., Wakshull E., Hancock W. S., *et al.*, (2006). Combination of Abundant Protein Depletion and Multi-Lectin Affinity Chromatography (M-LAC) for Plasma Protein Biomarker Discovery. *Journal of Proteome Research*, 6 (2), p. 662-671. DOI: 10.1021/pr060413k.

26. Lequin R. M., (2005). Enzyme Imlmunoassay (EIA)/Enzyme-Linked Immunosorbent Assay (ELISA). *Clinical Chemistry*, 51 (12), p. 2415-2418. DOI: 10.1373/clinchem.2005.051532.

27. Bantscheff M., Schirle M., Sweetman G., *et al.*, (2007). Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389 (4), p. 1017-1031. DOI: 10.1007/s00216-007-1486-6.

28. Dayon L., Turck N., Scherl A., *et al.*, (2010). From Relative to Absolute Quantification of Tryptic Peptides with Tandem Mass Tags: Application to Cerebrospinal Fluid. *Chimia*, 64 (3), p. 132-135. DOI: 10.2533/chimia.2010.132.

29. Zhu W., Smith J. W. & Huang C.-M., (2010). Mass Spectrometry-Based Label-Free Quantitative Proteomics. *Journal of Biomedicine and Biotechnology*, 2010, p. 1-7. DOI: 10.1155/2010/840518.

30. Lange V., Picotti P., Domon B., *et al.*, (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular Systems Biology*, 4 (1). DOI: 10.1038/msb.2008.61.

31. Laskowitz D. T., Kasner S. E., Saver J., *et al.*, (2009). Clinical Usefulness of a Biomarker-Based Diagnostic Test for Acute Stroke: The Biomarker Rapid Assessment in Ischemic Injury (BRAIN) Study. *Stroke*, 40 (1), p. 77-85. DOI: 10.1161/STROKEAHA.108.516377.

32. Sibon I., Rouanet F., Meissner W., *et al.*, (2009). Use of the Triage Stroke Panel in a neurologic emergency service. *American Journal of Emergency Medicine*, 27 (5), p. 558-562. DOI: 10.1016/j.ajem.2008.05.001.

33. Vanni S., Polidori G., Pepe G., *et al.*, (2011). Use of Biomarkers in Triage of Patients with Suspected Stroke. *The Journal of Emergency Medicine*, 40 (5), p. 499-505. DOI: 10.1016/j.jemermed.2008.09.028.

34. Knauer C., Knauer K., Muller S., *et al.*, (2012). A biochemical marker panel in MRI-proven hyperacute ischemic stroke - a prospective study. *BMC Neurology*, 12 (1), p. 14. DOI: 10.1186/1471-2377-12-14.

35. Fu Q., Zhu J. & Van Eyk J. E., (2010). Comparison of Multiplex Immunoassay Platforms. *Clinical Chemistry*, 56 (2), p. 314-318. DOI: 10.1373/clinchem.2009.135087.

36. Deeks J. J. & Altman D. G., (2004). Diagnostic tests 4: likelihood ratios. *British Medical Journal*, 329 (7458), p. 168-169. DOI: 10.1136/bmj.329.7458.168.

37. Naglie G., Krahn M. D., Naimark D., *et al.*, (1997). Primer on medical decision analysis: Part 3—Estimating probabilities and utilities. *Medical decision making*, 17 (2), p. 136-141.

38. Pletcher M. J. & Pignone M., (2011). Evaluating the Clinical Utility of a Biomarker A Review of Methods for Estimating Health Impact. *Circulation*, 123 (10), p. 1116-1124. DOI: 10.1161/CIRCULATIONAHA.110.943860.

39. Youden W. J., (1950). Index for rating diagnostic tests. *Cancer*, 3 (1), p. 32-35. DOI: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

40. Coffin M. & Sukhatme S., (1997). Receiver Operating Characteristic Studies and Measurement Errors. *Biometrics*, 53 (3), p. 823-837. DOI: 10.2307/2533545.

41. Perkins N. J. & Schisterman E. F., (2006). The Inconsistency of "Optimal" Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve. *American Journal of Epidemiology*, 163 (7), p. 670-675. DOI: 10.1093/aje/kwj063.

42. Fawcett T., (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27 (8), p. 861-874. DOI: 10.1016/j.patrec.2005.10.010.

43. McClish D. K., (1989). Analyzing a Portion of the ROC Curve. *Medical Decision Making*, 9 (3), p. 190-195. DOI: 10.1177/0272989X8900900307.

44. Altman D. G. & Bland J. M., (2009). Parametric v non-parametric methods for data analysis. *British Medical Journal*, 338, p. a3167. DOI: 10.1136/bmj.a3167.

45. Mann H. B. & Whitney D. R., (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18 (1), p. 50-60. DOI: 10.1214/aoms/1177730491.

46. Hanley J. A. & McNeil B. J., (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143 (1), p. 29-36.

47. Fisher R. A., (1922). On the Interpretation of χ2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85 (1), p. 87-94. DOI: 10.2307/2340521.

48. Agresti A., (1992). A Survey of Exact Inference for Contingency Tables. *Statistical Science*, 7 (1), p. 131-153. DOI: 10.1214/ss/1177011454.

49. Pepe M. S., (2003). *The statistical evaluation of medical tests for classification and prediction*, Oxford: Oxford University Press.

50. Hanley J. A., (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical Decision Making*, 8 (3), p. 197-203.

51. Hanley J. A. & McNeil B. J., (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148 (3), p. 839-843.

52. DeLong E. R., DeLong D. M. & Clarke-Pearson D. L., (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44 (3), p. 837-845.

53. Bandos A. I., Rockette H. E. & Gur D., (2005). A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Statistics in Medicine*, 24 (18), p. 2873-2893.

54. Braun T. M. & Alonzo T. A., (2008). A modified sign test for comparing paired ROC curves. *Biostat*, 9 (2), p. 364-372. DOI: 10.1093/biostatistics/kxm036.

55. Venkatraman E. S. & Begg C. B., (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, 83 (4), p. 835-848. DOI: 10.1093/biomet/83.4.835.

56. Venkatraman E. S., (2000). A Permutation Test to Compare Receiver Operating Characteristic Curves. *Biometrics*, 56 (4), p. 1134-1138. DOI: 10.1111/j.0006-341X.2000.01134.x.

57. Robin X., Turck N., Hainard A., *et al.*, (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, p. 77. DOI: 10.1186/1471-2105-12-77.

58. Anon, (1998). Consensus Report of the Working Group on: "Molecular and Biochemical Markers of Alzheimer's Disease." *Neurobiology of Aging*, 19 (2), p. 109-116. DOI: 10.1016/S0197-4580(98)00022-0.

59. Hastie T., Tibshirani R. & Friedman J., (2003). *Elements of Statistical Learning: data mining, inference, and prediction* Springer-Verlag., New York.

60. Saeys Y., Inza I. & Larranaga P., (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23 (19), p. 2507-2517. DOI: 10.1093/bioinformatics/btm344.

61. Hilario M. & Kalousis A., (2008). Approaches to dimensionality reduction in proteomic biomarker studies. *Briefings in Bioinformatics*, 9 (2), p. 102-118. DOI: 10.1093/bib/bbn005.

62. Dziuda D. M., (2010). *Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data*, John Wiley & Sons.

63. Baggerly K. A., Morris J. S., Wang J., *et al.*, (2003). A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *PROTEOMICS*, 3 (9), p. 1667-1672. DOI: 10.1002/pmic.200300522.

64. Petricoin E. F., Ardekani A. M., Hitt B. A., *et al.*, (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359 (9306), p. 572-577. DOI: 10.1016/S0140-6736(02)07746-2.

65. Peng S., Xu Q., Ling X. B., *et al.*, (2003). Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letters*, 555 (2), p. 358-362. DOI: 10.1016/S0014-5793(03)01275-4.

66. Cima I., Schiess R., Wild P., *et al.*, (2011). Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. *Proceedings of the National Academy of Sciences*, 108 (8), p. 3342-3347. DOI: 10.1073/pnas.1013699108.

67. Ahn H. S., Shin Y. S., Park P. J., *et al.*, (2012). Serum biomarker panels for the diagnosis of gastric adenocarcinoma. *British Journal of Cancer*, 106 (4), p. 733-739. DOI: 10.1038/bjc.2011.592.

68. Hill M. D., Jackowski G., Bayer N., *et al.*, (2000). Biochemical markers in acute ischemic stroke. *Canadian Medical Association Journal*, 162 (8), p. 1139-1140.

69. Faca V. M., Song K. S., Wang H., *et al.*, (2008). A Mouse to Human Search for Plasma Proteome Changes Associated with Pancreatic Tumor Development. *PLoS Medicine*, 5 (6), p. e123. DOI: 10.1371/journal.pmed.0050123.

70. Montaner J., Perea-Gainza M., Delgado P., *et al.*, (2008). Etiologic Diagnosis of Ischemic Stroke Subtypes With Plasma Biomarkers. *Stroke*, 39 (8), p. 2280-2287. DOI: 10.1161/STROKEAHA.107.505354.

71. Lejon V., Roger I., Mumba Ngoyi D., *et al.*, (2008). Novel Markers for Treatment Outcome in Late-Stage Trypanosoma brucei gambiense Trypanosomiasis. *Clinical Infectious Diseases*, 47 (1), p. 15-22. DOI: 10.1086/588668.

72. Reynolds M. A., Kirchick H. J., Dahlen J. R., *et al.*, (2003). Early Biomarkers of Stroke. *Clinical Chemistry*, 49 (10), p. 1733-1739. DOI: 10.1373/49.10.1733.

73. Patz E. F., Campa M. J., Gottlin E. B., *et al.*, (2007). Panel of Serum Biomarkers for the Diagnosis of Lung Cancer. *Journal of Clinical Oncology*, 25 (35), p. 5578-5583. DOI: 10.1200/JCO.2007.13.5392.

74. Seeber B., Sammel M. D., Fan X., *et al.*, (2008). Panel of markers can accurately predict endometriosis in a subset of patients. *Fertility and Sterility*, 89 (5), p. 1073-1081. DOI: 10.1016/j.fertnstert.2007.05.014.

75. Frenzel J., Gessner C., Sandvoss T., *et al.*, (2011). Outcome Prediction in Pneumonia Induced ALI/ARDS by Clinical Features and Peptide Patterns of BALF Determined by Mass Spectrometry. *PLoS ONE*, 6 (10), p. e25544. DOI: 10.1371/journal.pone.0025544.

76. Reddy A., Wang H., Yu H., *et al.*, (2008). Logical Analysis of Data (LAD) model for the early diagnosis of acute ischemic stroke. *BMC Medical Informatics and Decision Making*, 8 (1), p. 30. DOI: 10.1186/1472-6947-8-30.

77. Torkaman A., Charkari N. M. & Aghaeipour M., (2011). An approach for leukemia classification based on cooperative game theory. *Analytical Cellular Pathology*, 34 (5), p. 235-246. DOI: 10.3233/ACP-2011-0016.

78. Prados J., Kalousis A., Sanchez J.-C., *et al.*, (2004). Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics*, 4 (8), p. 2320-2332. DOI: 10.1002/pmic.200400857.

79. Oh J. H., Kim Y. B., Gurnani P., *et al.*, (2008). Biomarker Selection and Sample Prediction for Multi-Category Disease on MALDI-TOF Data. *Bioinformatics*, 24 (16), p. 1812-1818. DOI: 10.1093/bioinformatics/btn316.

80. Farlow E. C., Vercillo M. S., Coon J. S., *et al.*, (2010). A multi-analyte serum test for the detection of non-small cell lung cancer. *British Journal of Cancer*, 103 (8), p. 1221-1228. DOI: 10.1038/sj.bjc.6605865.

81. Thiel S. W., Rosini J. M., Shannon W., *et al.*, (2010). Early prediction of septic shock in hospitalized patients. *Journal of Hospital Medicine*, 5 (1), p. 19-25. DOI: 10.1002/jhm.530.

82. Wang P., Kim Y., Pollack J., *et al.*, (2004). Boosted PRIM with Application to Searching for Oncogenic Pathway of Lung Cancer. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*. IEEE Computer Society, pp. 604-609.

83. Dyson G., Frikke-Schmidt R., Nordestgaard B. G., *et al.*, (2007). An application of the patient rule-induction method for evaluating the contribution of the Apolipoprotein E and Lipoprotein Lipase genes to predicting ischemic heart disease. *Genetic Epidemiology*, 31 (6), p. 515-527. DOI: 10.1002/gepi.20225.

84. Nannings B., Abu-Hanna A. & de Jonge E., (2008). Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. *International Journal of Medical Informatics*, 77 (4), p. 272-279. DOI: 10.1016/j.ijmedinf.2007.06.007.

85. Frikke-Schmidt R., Tybjærg-Hansen A., Schnohr P., *et al.*, (2010). Common clinical practice versus new PRIM score in predicting coronary heart disease risk. *Atherosclerosis*, 213 (2), p. 532-538. DOI: 10.1016/j.atherosclerosis.2010.07.028.

86. Warner J. H., Liang Q., Sarkar M., *et al.*, (2010). Adaptive regression modeling of biomarkers of potential harm in a population of U.S. adult cigarette smokers and nonsmokers. *BMC Medical Research Methodology*, 10 (1), p. 19. DOI: 10.1186/1471-2288-10-19.

87. Brasier A. R., Garcia J., Wiktorowicz J. E., *et al.*, (2012). Discovery Proteomics and Nonparametric Modeling Pipeline in the Development of a Candidate Biomarker Panel for Dengue Hemorrhagic Fever. *Clinical and Translational Science*, 5 (1), p. 8-20. DOI: 10.1111/j.1752-8062.2011.00377.x.

88. Breiman L., (2001). Random Forests. *Machine Learning*, 45 (1), p. 5-32. DOI: 10.1023/A:1010933404324.

89. Brasier A. R., Ju H., Garcia J., *et al.*, (2012). A Three-Component Biomarker Panel for Prediction of Dengue Hemorrhagic Fever. *The American Journal of Tropical Medicine and Hygiene*, 86 (2), p. 341-348. DOI: 10.4269/ajtmh.2012.11-0469.

90. Lundgren D. H., Hwang S.-I., Wu L., *et al.*, (2010). Role of spectral counting in quantitative proteomics. *Expert Review of Proteomics*, 7 (1), p. 39-53. DOI: 10.1586/epr.09.69.

91. Yuan Z. & Ghosh D., (2008). Combining Multiple Biomarker Models in Logistic Regression. *Biometrics*, 64 (2), p. 431-439. DOI: 10.1111/j.1541-0420.2007.00904.x.

92. Gomar J. J., Bobes-Bascaran M. T., Conejero-Goldberg C., *et al.*, (2011). Utility of Combinations of Biomarkers, Cognitive Markers, and Risk Factors to Predict Conversion From Mild Cognitive Impairment to Alzheimer Disease in Patients in the Alzheimer's Disease Neuroimaging Initiative. *Archives of General Psychiatry*, 68 (9), p. 961-969. DOI: 10.1001/archgenpsychiatry.2011.96.

93. Visintin I., Feng Z., Longton G., *et al.*, (2008). Diagnostic Markers for Early Detection of Ovarian Cancer. *Clinical Cancer Research*, 14 (4), p. 1065-1072. DOI: 10.1158/1078-0432.CCR-07-1569.

94. Welsh P., Barber M., Langhorne P., *et al.*, (2009). Associations of inflammatory and haemostatic biomarkers with poor outcome in acute ischaemic stroke. *Cerebrovascular Diseases*, 27 (3), p. 247-253. DOI: 10.1159/000196823.

95. Wicki J., Perneger T. V., Junod A. F., *et al.*, (2001). Assessing Clinical Probability of Pulmonary Embolism in the Emergency Ward: A Simple Score. *Archives of Internal Medicine*, 161 (1), p. 92-97. DOI: 10.1001/archinte.161.1.92.

96. Nolen B. M., Langmead C. J., Choi S., *et al.*, (2011). Serum biomarker profiles as diagnostic tools in lung cancer. *Cancer biomarkers: section A of Disease markers*, 10 (1), p. 3-12. DOI: 10.3233/CBM-2012-0229.

97. Friedman J., (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, 28 (2), p. 337-407. DOI: 10.1214/aos/1016218223.

98. Yurkovetsky Z., Skates S., Lomakin A., *et al.*, (2010). Development of a Multimarker Assay for Early Detection of Ovarian Cancer. *Journal of Clinical Oncology*, 28 (13), p. 2159-2166. DOI: 10.1200/JCO.2008.19.2484.

99. Liu J. J., Cutler G., Li W., *et al.*, (2005). Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics*, 21 (11), p. 2691-2697. DOI: 10.1093/bioinformatics/bti419.

100. Schramm A., Schulte J. H., Klein-Hitpass L., *et al.*, (2005). Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene*, 24 (53), p. 7902-7912.

101. Zervakis M., Blazadonakis M. E., Tsiliki G., *et al.*, (2009). Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC Bioinformatics*, 10, p. 53. DOI: 10.1186/1471-2105-10-53.

102. Gomes A. L. V., Wee L. J. K., Khan A. M., *et al.*, (2010). Classification of Dengue Fever Patients Based on Gene Expression Data Using Support Vector Machines. *PLoS ONE*, 5 (6), p. e11267. DOI: 10.1371/journal.pone.0011267.

103. Wu B., Abbott T., Fishman D., *et al.*, (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19 (13), p. 1636-1643. DOI: 10.1093/bioinformatics/btg210.

104. Ressom H. W., Varghese R. S., Drake S. K., *et al.*, (2007). Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics*, 23 (5), p. 619-626. DOI: 10.1093/bioinformatics/btl678.

105. Barla A., Jurman G., Riccadonna S., *et al.*, (2008). Machine learning methods for predictive proteomics. *Briefings in Bioinformatics*, 9 (2), p. 119-128. DOI: 10.1093/bib/bbn008.

106. Guan W., Zhou M., Hampton C. Y., *et al.*, (2009). Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics*, 10 (1), p. 259. DOI: 10.1186/1471-2105-10-259.

107. McDonald R. A., Skipp P., Bennell J., *et al.*, (2009). Mining whole-sample mass spectrometry proteomics data for biomarkers - An overview. *Expert Systems with Applications*, 36 (3, Part 1), p. 5333-5340. DOI: 10.1016/j.eswa.2008.06.133.

108. Wild N., Karl J., Grunert V. P., *et al.*, (2008). Diagnosis of rheumatoid arthritis: multivariate analysis of biomarkers. *Biomarkers*, 13 (1), p. 88-105. DOI: 10.1080/13547500701669410.

109. Burges C. J. C., (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2 (2), p. 121-167. DOI: 10.1023/A:1009715923555.

110. Varma S. & Simon R., (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, p. 91. DOI: 10.1186/1471-2105-7-91.

111. Tibshirani R. J., (2009). A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, 3 (2), p. 822-829. DOI: 10.1214/08-AOAS224.

112. Zhang Z., Yu Y., Xu F., *et al.*, (2007). Combining multiple serum tumor markers improves detection of stage I epithelial ovarian cancer. *Gynecologic Oncology*, 107 (3), p. 526-531. DOI: 10.1016/j.ygyno.2007.08.009.

113. Donach M., Yu Y., Artioli G., *et al.*, (2010). Combined use of biomarkers for detection of ovarian cancer in high-risk women. *Tumor Biology*, 31 (3), p. 209-215. DOI: 10.1007/s13277-010-0032-x.

114. Flores-Fernández J. M., Herrera-López E. J., Sánchez-Llamas F., *et al.*, (2012). Development of an optimized multi-biomarker panel for the detection of lung cancer based on principal component analysis and artificial neural network modeling. *Expert Systems with Applications*, 39 (12), p. 10851-10856. DOI: 10.1016/j.eswa.2012.03.008.

115. Webb G. I., Boughton J. R. & Wang Z., (2005). Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning*, 58 (1), p. 5-24. DOI: 10.1007/s10994-005-4258-6.

116. Ralhan R., DeSouza L. V., Matta A., *et al.*, (2008). Discovery and Verification of Head-and-neck Cancer Biomarkers by Differential Protein Expression Analysis Using iTRAQ Labeling, Multidimensional Liquid Chromatography, and Tandem Mass Spectrometry. *Molecular & Cellular Proteomics*, 7 (6), p. 1162-1173. DOI: 10.1074/mcp.M700500-MCP200.

117. Ring B. Z., Seitz R. S., Beck R., *et al.*, (2006). Novel Prognostic Immunohistochemical Biomarker Panel for Estrogen Receptor-Positive Breast Cancer. *J Clin Oncol*, 24 (19), p. 3039-3047. DOI: 10.1200/JCO.2006.05.6564.

118. Damman P., Beijk M. A. M., Kuijt W. J., *et al.*, (2011). Multiple Biomarkers at Admission Significantly Improve the Prediction of Mortality in Patients Undergoing Primary Percutaneous Coronary Intervention for Acute ST-Segment Elevation Myocardial Infarction. *Journal of the American College of Cardiology*, 57 (1), p. 29-36. DOI: 10.1016/j.jacc.2010.06.053.

119. Knickerbocker T., Chen J. R., Thadhani R., *et al.*, (2007). An integrated approach to prognosis using protein microarrays and nonparametric methods. *Molecular Systems Biology*, 3 (123), p. 1-8. DOI: 10.1038/msb4100167.

120. Whiteley W., Tseng M.-C. & Sandercock P., (2008). Blood Biomarkers in the Diagnosis of Ischemic Stroke: A Systematic Review. *Stroke*, 39 (10), p. 2902-2909. DOI: 10.1161/STROKEAHA.107.511261.

121. Feng Z. & Yasui Y., (2004). Statistical considerations in combining biomarkers. *Disease Markers*, 20 (2), p. 45-51.

122. Hesterberg T., Moore D. S., Monaghan S., *et al.*, (2005). Bootstrap Methods and Permutation Tests. In *Introduction to the Practice of Statistics*. p. 896.

123. Carpenter J. & Bithell J., (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in medicine*, 19 (9), p. 1141-1164. DOI: 10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F.

124. Bhaskar H., Hoyle D. C. & Singh S., (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in Biology and Medicine*, 36 (10), p. 1104-1125.

125. Baggerly K. A., Morris J. S. & Coombes K. R., (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20 (5), p. 777-785. DOI: 10.1093/bioinformatics/btg484.

2

Bioinformatics for protein biomarker panel classification: What is needed to bring biomarker panels into in vitro diagnostics?

This review summarizes the current usage of combinations of biomarkers for clinical purposes, resulting from the increasing number of protein biomarkers discovered by proteomics techniques. Most of these biomarkers display low sensitivities or specificities, preventing their translation into *in vitro* diagnostics and, consequently, to clinical practice. This observation led to the idea that combining these proteins into panels of biomarkers could improve their clinical usefulness, which represents the main hypothesis of this thesis.

This chapter presents the state-of-the-art in this field. It summarizes the main methods employed, starting with very basic statistic to advanced machine learning methods, and outlines their strength and weaknesses. It discusses the challenges raised by this kind of analysis, namely over-fitting and validation issues resulting from the high dimensionality of the datasets, the low number of patients and poor reproducibility of proteomics techniques.

I fully wrote this review with help and input from the co-authors.

EXPERT REVIEWS

Bioinformatics for protein biomarker panel classification: what is needed to bring biomarker panels into *in vitro* diagnostics?

Expert Rev. Proteomics 6(6), 675–689 (2009)

Xavier Robin, Natacha Turck, Alexandre Hainard, Frédérique Lisacek, Jean-Charles Sanchez† and Markus Müller

†*Author for correspondence*
Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
jean-charles.sanchez@unige.ch

A large number of biomarkers have been discovered over the past few years using proteomics techniques. Unfortunately, most of them are neither specific nor sensitive enough to be translated into *in vitro* diagnostics and routine clinical practice. From this observation, the idea of combining several markers into panels to improve clinical performances has emerged. In this article, we present a discussion of the bioinformatics aspects of biomarker panels and the concomitant challenges, including high dimensionality and low patient number and reproducibility.

KEYWORDS: biomarker • classification • combination of biomarkers • *in vitro* diagnostics • panel • proteomics • validation

Background

As part of clinical practice, it is common to measure the concentration of a protein, known as a biomarker, in a biological sample to diagnose a disease, predict the outcome early or monitor a therapy. Examples of commonly accepted biomarkers include troponin I for detecting acute myocardial infarction, prostate-specific antigen for the screening of prostate cancer, glycated hemoglobin for the control of long-term glycemia and C-reactive protein for assessment of inflammation. Proteomics techniques, such as 2DGE [1,2] and mass spectrometry [2–5], have led to the discovery of numerous biomarkers, most of which are not currently available to medical practitioners. Possible explanations for this gap between proteomics research and routine practice are technical (e.g., the time and huge costs required to validate these molecules, as well as the accuracy of assays not being high enough to be translated directly into clinical practice) and biological (e.g., inter- and intra-individual variability).

When several biomarkers are measured, they are often considered separately, irrespective of the additional information contained in their joined interpretation. Combining several biomarkers into a single classification rule helps to improve their classification accuracy and, therefore, their clinical usefulness. Hereafter, we will call such a combination a panel. Potentially, a panel could even combine clinical parameters, such as age, sex, physiological constants or clinical scores, with biomarkers [6]. Similar to a single marker, a panel allows us to answer different clinical questions. Apart from increasing accuracy, biomarker panels help in the study of different pathophysiological pathways and shed light on diseases from different angles. For instance, in the context of a brain damage condition (e.g., aneurysmal subarachnoid hemorrhage), Turck *et al.* recently demonstrated that a combination of brain parameters associated with a clinical score and a cardiac biomarker could predict 6-month outcomes better than the biomarkers taken individually could [7]. In the same manner, Hainard *et al.* proposed a combination of inflammatory cytokines and one brain damage marker [8]. In both cases, the combination of different kinds of biomarkers improved the classification.

In contrast to the traditional single-analyte interpretation, several new challenges arise, which could also explain why panels are not yet

 10.1586/EPR.09.83 ISSN 1478-9450

widespread. First, appropriate methods are required to combine information from multiple biomarkers. These methods must be efficient and yield correct patient classification, but they must also be comprehensible to medical practitioners in order to gain acceptance. Second, the risk of overfitting the data is increased because of the higher dimensionality [9–11]. A careful validation is required to ensure that a panel truly performs better than individual biomarkers, hence avoiding raising false hopes. Finally, appropriate experimental design [12,13] and validation are the most important factors for ensuring the quality of the results.

After a short overview of *in vitro* diagnostics (IVD), we will review recent papers that describe combinations of biomarkers and/or clinical parameters in panels, and examine whether they addressed these new challenges and, if so, how. We mainly focus on protein biomarker panels but also include related work on analyzing protein or gene expression microarray and protein mass spectrometry data where we deem it relevant for protein panels. We will also review methods that allow the validation of obtained models and their performance, as well as the strategies available to compare different panels and their combinations. This review addresses clinical researchers who seek a basic introduction to statistical methods and the pitfalls of biomarker-panel research and statisticians who would like to learn more about recent work and the clinical aspects of this subject.

From discovery to IVD

In vitro diagnostics encompasses any type of assay performed on a patient sample in a controlled environment to answer a clinical question, including diagnostic, prognostic or monitoring tests. It typically includes point-of-care tests, which are quick and simple assays performed beside the patient with portative equipment, and laboratory tests, which are performed by trained personnel in dedicated clinical chemistry laboratories.

Vitzthum *et al.* reviewed the need in proteomics to push discovered molecules into IVD [14]. The crucial points are that the classification must be reliable and provide information that is valuable for decision making; measurements must be both exact and robust, and the test accuracy must meet sufficient (positive or negative) predictive values.

The target performance of IVD tests must be chosen according to the clinical question. As pointed out by Dodd and Pepe, "large monetary costs result from high false-positive rates" [15]. Similarly, failure to diagnose a disease can dramatically impact on a patient's health, which may even lead to death. Therefore, IVD (single biomarker or panel) as a helpful clinical practice must display sufficient discriminative power and answer a well-defined question. In other words, one should focus on high sensitivity and/or specificity or high predictive values rather than global accuracy.

An IVD test aims to determine the state of the patient. For biomarker tests, a decision threshold (also known as a cut-off) is usually chosen. Any value below the cut-off will indicate that the test result is negative, while a value above the threshold will be deemed as a positive result. The test result, together with the observed true outcome, will define the sensitivity and specificity (see Table 1 for definitions).

Predictive tests can be split into two categories: 'rule-out' and 'rule-in' tests. Rule-out tests reject negative patients while avoiding false negatives. In these tests, the sensitivity is of prime importance, as is the negative predictive value. However, the level of false positive must be kept low enough in order to preserve both specificity and positive predictive value at acceptable levels. When the test is applied to asymptomatic patients, it is termed a screening test. A negative result to a screening test implies that the patient is highly likely to be healthy, while a positive result only means that more investigations are required. For example, in the context of human African trypanosomiasis (HAT), a potential rule-out test would be applied to exclude the patients not infected by the parasite. All patients with a negative test would then be classed as free of the parasite, with a very high confidence. Similarly, rule-in tests (also called confirmatory tests) try to include only positive patients and generate as few false positives as possible. The specificity and positive predictive values must be very high. A rule-in test applied in the HAT field would select only patients with parasites in the brain (stage 2 of the disease), who would be subsequently subjected to a very toxic treatment. Patients without brain infection (stage 1) must be excluded, because they could potentially be killed by the inappropriate medication [8].

Predictive values (negative or positive) need to take the class prevalence into account since even a test with a very high specificity could have a low positive predictive value. If the prevalence of the disease is very low, there would be a larger number of false positive, only because of the larger number of controls. This property makes predictive values more difficult to compute than specificity or sensitivity. Despite this complication, predictive values are usually more valuable because they express the probability of the patient being truly positive or negative for a given group of patients.

Commercial panels

From a commercial point of view, McCormick showed how both pharmaceutical companies and medical practitioners could profit from biomarkers and biomarker panels to predict the safety of a treatment, identify risk and responder candidates, and monitor therapies [101]. However, they pointed out that the acceptance of biomarkers is hindered by the lack of data sharing (owing to technical or strategic reasons), as well as insufficient validation and targeting.

In the USA, medical devices (including IVD) must obtain approval by the US FDA. Hackett and Gutman highlighted the difficulties that are raised by the combination of several markers and the use of statistical models [16]. FDA review procedures for device acceptance focus on the test result, and a simple model can be accepted at the condition that it is independently validated.

To our knowledge, only the Biosite® company sells panels of protein biomarkers for blood samples. The Triage® Stroke Panel simultaneously measures four markers (namely matrix metalloproteinase 9, brain natriuretic peptide, D-dimer and S100β) and computes a multimarker index using a proprietary algorithm. Two cut-offs are defined, associated with a high or low risk for the patient having a stroke, while patients in the intermediate region require further investigation. It was accepted by the FDA for premarket

Table 1. Clinical classification definitions.

Word	Common abbreviation	Formula	Definition
Prevalence			Frequency of the positive occurrence in the studied population
Rule in (confirmatory)			A test performed in an attempt to confirm the presence of a disease
Rule out (screening)			A test performed in an attempt to exclude the presence of a disease
True negatives	TN		Negative patients correctly classified as negatives
True positives	TP		Positive patients correctly classified as positives
False negatives	FN		Positive patients incorrectly classified as negatives
False positives	FP		Negative patients incorrectly classified as positives
Sensitivity	SE	TP/(TP+FN)	Proportion of positive patients correctly detected by the test
Specificity	SP	TN/(TN+FP)	Proportion of negative patients correctly rejected by the test
Positive predictive value	PPV	TP/(TP+FP)	Proportion of positive tests that correctly indicate positive patients
Negative predictive value	NPV	TN/(TN+FN)	Proportion of negative tests that correctly indicate negative patients
Odds ratio	OR	(SE/(1-SE))(SP/(1-SP))	Effect of a given increase of the studied marker

approval application in 2005, was withdrawn by the manufacturer 1 year later to allow further clinical studies, but it was recently reintroduced. The Triage Stroke Panel was applied by Vanni *et al.* in a neurological emergency service to discriminate patients with or without stroke among those suspected of having stroke [17]. Sibon *et al.* compared it with an established neurological scale evaluated by nurses [18]. Brouns *et al.* analyzed only the D-dimer measurement to compare it with the assay of another manufacturer to discriminate small-artery and large-artery acute ischemic stroke, but they did not make use of the multimarker index score [19].

The same company previously marketed in 1999 a Triage® Cardiac Panel for the diagnosis of cardiovascular diseases. This test measures three proteins, known as cardiac markers (namely creatine kinase MB, myoglobin and troponin I) [20]. However, it cannot truly be called a panel as the measurements are not combined into a single final score.

Applied Genomics sells several immunohistochemistry panels. Tissue arrays are stained, and each staining is assessed in a binary manner. The results are then combined with a Cox proportional hazards model into a single score stratifying patients into low, medium or high risk. One of the available panels provides prognostic information for breast cancer outcome [18].

Tools for panels

History

In terms of biomarkers, a panel is the combination of more than one variable into a single classification rule. The idea of combining several medical parameters to obtain an improved patient classification is not new. In psychiatry, Hoffer and Osmond applied a combination of neuropsychiatric variables in the early 1960s to distinguish schizophrenic patients from normal individuals [21]. They defined 145 questions that could be answered by true or false, covering perceptions, thoughts and feelings. Complex algorithms would then compute several scores. However, the set of questions and the scoring algorithms were not justified. Later, in 1988, the World Federation of Neurological Surgeons (WFNS) score was defined to assess patients' neurological status [22]. It consists of the combination of three easy-to-assess clinical variables. Eye, verbal and motor responses are evaluated on a scale ranging from one to four, five and six, respectively. An intermediate score ranging from three to 15 is computed, and the final score depends on the range of this intermediate score and the presence of a motor deficit.

In the field of biomarkers, Woolas *et al.* showed the potential of using several serum markers together in 1993 [23]. They observed that most of their patients with stage 1 ovarian cancer were positive for at least one of the three markers they tested. However, they did not use this observation to make a true statistical combination. In 2000, Hill *et al.* were among the first to report the use of a panel of protein biomarkers [24]. They tested four biomarkers, and they observed that 93% of their acute ischemic stroke patients were positive for at least one of the four markers of the panel.

As detailed in the later section 'Classification using panels', panels can also combine biomarkers and clinical parameters. However, prior to discussing the various approaches for panel classification, we briefly review some important data preprocessing and data-normalization steps, which are performed prior to classification.

Preprocessing

Normalization & reproducibility

Several types of errors can disturb the results of biomarker concentration measurements and mitigate reproducibility. It has been shown that sample collection from different centers and by different nurses as well as sample handling variability (i.e., sample container, time to freezing and storage temperature) and instrumental errors can lead to measurement variations [25,26]. When dealing with high-dimensional mass spectra, reproducibility of the experiments becomes a problem, and it has been shown that proper sample and data processing, as well as feature selection,

are of major importance [9]. Furthermore, biological variability between different patients owing to sex, age, treatment, lifestyle and chronic diseases, or even within a single patient taken at different times, can confuse the analysis. All these sources of variation make it difficult to compare the results of different experiments and to draw conclusions.

On the experimental side, normalization methods often require a 'calibration' sample, which has constant values over all the experiments [27]. Using calibration curves, concentration measurements of biomarkers can be adjusted for each patient and systematic offsets in the measurements reduced. However, only instrumental offsets can be reduced in this way and other offsets due to sample acquisition and treatment need further bioinformatics normalization.

This computational normalization equalizes the mean and variance of distributions of different biomarker measurements, making them more comparable. A very simple normalization method consists of the z-score transformation, which sets the mean to zero and the variance to one, but otherwise does not affect the shape of the distribution. Yeo *et al.* proposed the box–cox transformation family, which includes the logarithmic transformation, to obtain distributions closer to the normal one [28]. Another normalization method is the quantile normalization technique, where all values are transformed into their corresponding normal quantiles [29]. However, this is an extreme normalization and the structure of the data can be lost in the process. Based on technical and biological replicates, analysis of variance can calculate the bias and variance introduced by each processing step and lead to more accurate comparisons [13].

Feature selection

Another important preprocessing step is feature selection, which is crucial in high-dimensionality problems, such as mass spectra or microarrays, but is less important for lower dimensional biomarker panels. It consists of selecting the biomarkers and patient parameters that will be included in the panel. The choice of the feature-selection method strongly depends on the classification algorithm and the data [30]. It is also important to note that data for feature selection must not include the test data; otherwise, the test performance would be too optimistic. Saeys *et al.* classified the feature-selection methods into three categories: filter methods, wrapper methods and embedded methods [31]. Filter methods consider only the intrinsic properties of single features independently from classification. Conversely, wrapper and embedded methods perform the feature search at the same time as the classifier model is trained. In wrapper methods, the search for optimal features is performed by an optimization procedure, which evaluates the performance of a given classifier on different feature subsets. Embedded techniques can include or eliminate features during the classifier-training procedure. Such embedded techniques can be implemented, for example, in logistic regression, random forests, neural networks or support vector machines (SVMs; see later).

Several examples of feature selection are reported by Hilario and Kalousis [30]. Baggerly *et al.* used preprocessing and exhaustive search and genetic algorithms to reduce an initial 60 831 *m/z* value from mass spectrometry to filter 506 and then sets of one to five features, and then applied the feature sets to linear discriminant analysis [32]. Petricoin *et al.* also employed genetic algorithm with mass spectrometry, but in a wrapper method around a self-organizing map algorithm [33].

Classification using panels

Biomarker panels rely on a well-established field of statistics, known as multivariate classification or supervised learning. There is a vast amount of literature available, and much of it is summarized in excellent textbooks, such as that by Hastie *et al.* [34]. The classification task consists of attributing a class label to every patient by means of the vector of biomarker concentrations and clinical scores. In the case of two classes, this corresponds to dividing the space of all possible panel vectors into two distinct regions, one region for every class (Figure 1). The way the classifier determines these regions depends on the method used. In all cases, the algorithms learn these boundaries from training data, that is, a set of panel vectors known to belong to a diseased or healthy patient. Once the region boundaries are fixed, the performance can be evaluated on equally annotated but disjointed test data.

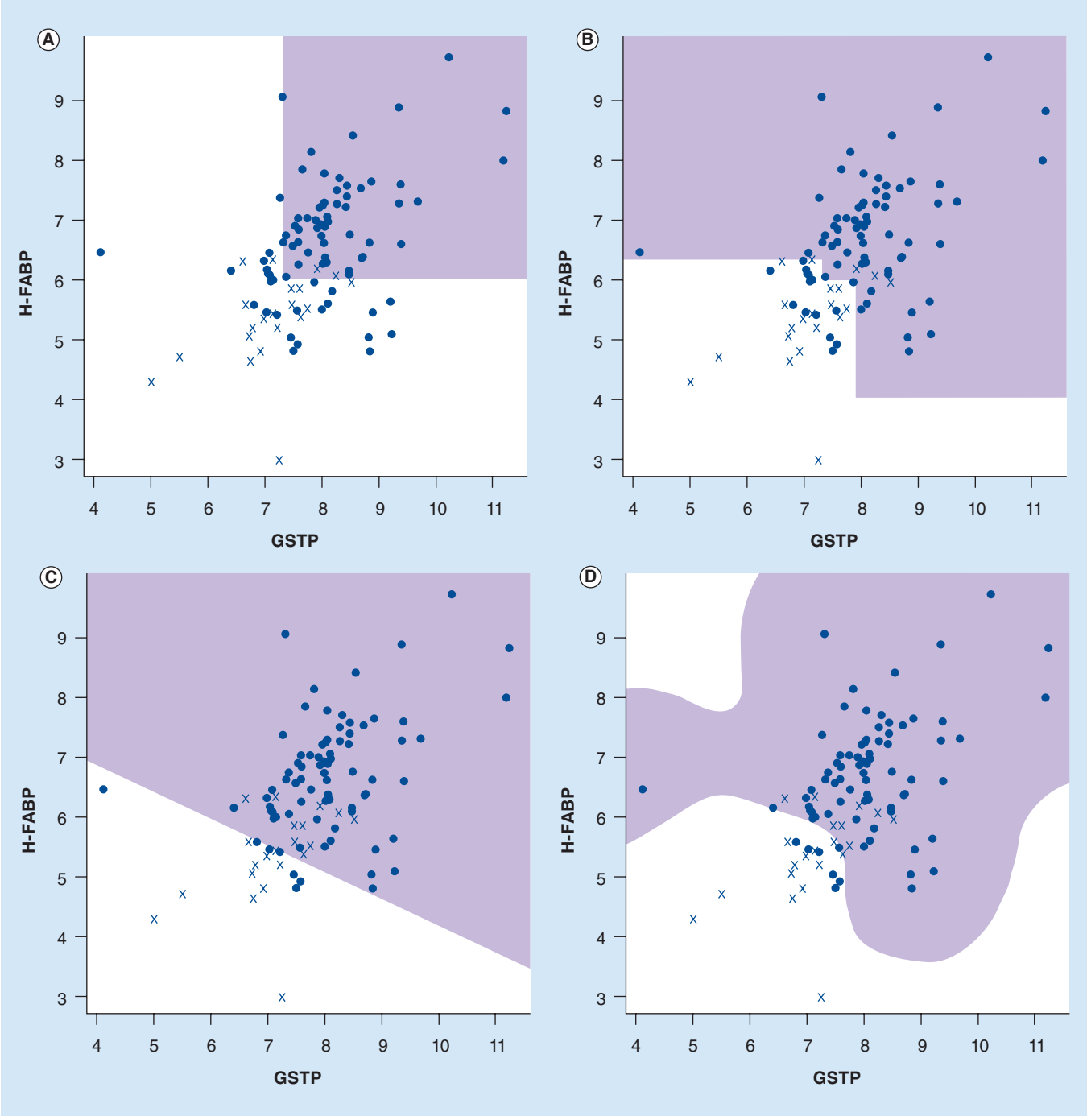
This approach may seem fairly straightforward, but two main problems must be dealt with: the low number of samples in the training set and overfitting the data. The former problem is paramount in many biomarker projects, since the number of patients is usually small (from a few to several hundred patients) compared with the number of markers. The patients are then only sparsely distributed in the panel vector space, and many parts of the class regions are only represented poorly or not at all in the training set, which makes it more difficult for the classifier to find the correct regions. Figure 1 illustrates this problem since neither the upper left nor the lower right corners contain any data points and, considering only these training data, it is impossible to predict the classifier results in these regions. The latter problem is perhaps less severe but equally important. Since the shape and smoothness of the boundaries between the class regions is not known (linear or curved), the regions obtained from the training data might be wrong even if they fit the training data very well, because the model defined in the classifier is wrong (i.e., the classifier might yield an arbitrarily curved boundary that is actually linear Figure 2). However, cross validation provides a means to at least partially mitigate this problem (see later). As a rule of thumb, the fewer patients there are in the training and test sets, the simpler the class boundaries should be to avoid overfitting, even if these simple boundaries cannot reproduce the true boundaries correctly.

We now discuss the main methods applied to define biomarker panels. Threshold-based methods and logistic regression are probably the most popular ones. Tree-based methods are also widely used, whereas SVM is a method of choice for many high-dimensional problems. We will now detail some methods and show how they are applied.

Threshold-based

In threshold-based methods, a set of thresholds, one for each biomarker, is selected, usually in a univariate manner (Figure 1A) [7,8,24,35–38]. Any value of a molecule below its

respective threshold will indicate that the test result is negative, while a value above the threshold will be deemed a positive result. In some rare cases, it can be necessary to reverse the order and to consider values below the threshold as positive results. The score of the test for a patient corresponds to the number of biomarker molecules, whose concentration value exceeds (or is below for negative biomarkers) the threshold. Similar to a majority voting, a patient is classified positively if this score is higher than



Figures 1. Classification by different methods. (GSTP and H-FABP concentrations illustrated in log scale). The gray area shows the region where the test would be considered positive by the method. Crosses and dots represent Stage 1 and Stage 2 human African trypanosomiasis patients, respectively. **(A)** Threshold-based methods split the space into boxes. **(B)** Decision trees can create more boxes. **(C)** Logistic regression divides the data with a straight line. **(D)** Support vector machines can compute complex separations but can also create linear partitions similarly to logistic regression (see Figure 4).
GSTP: Glutathione *S*-transferase Pi; H-FABP: Heart-type fatty acid-binding protein.
Redrawn from from [8].

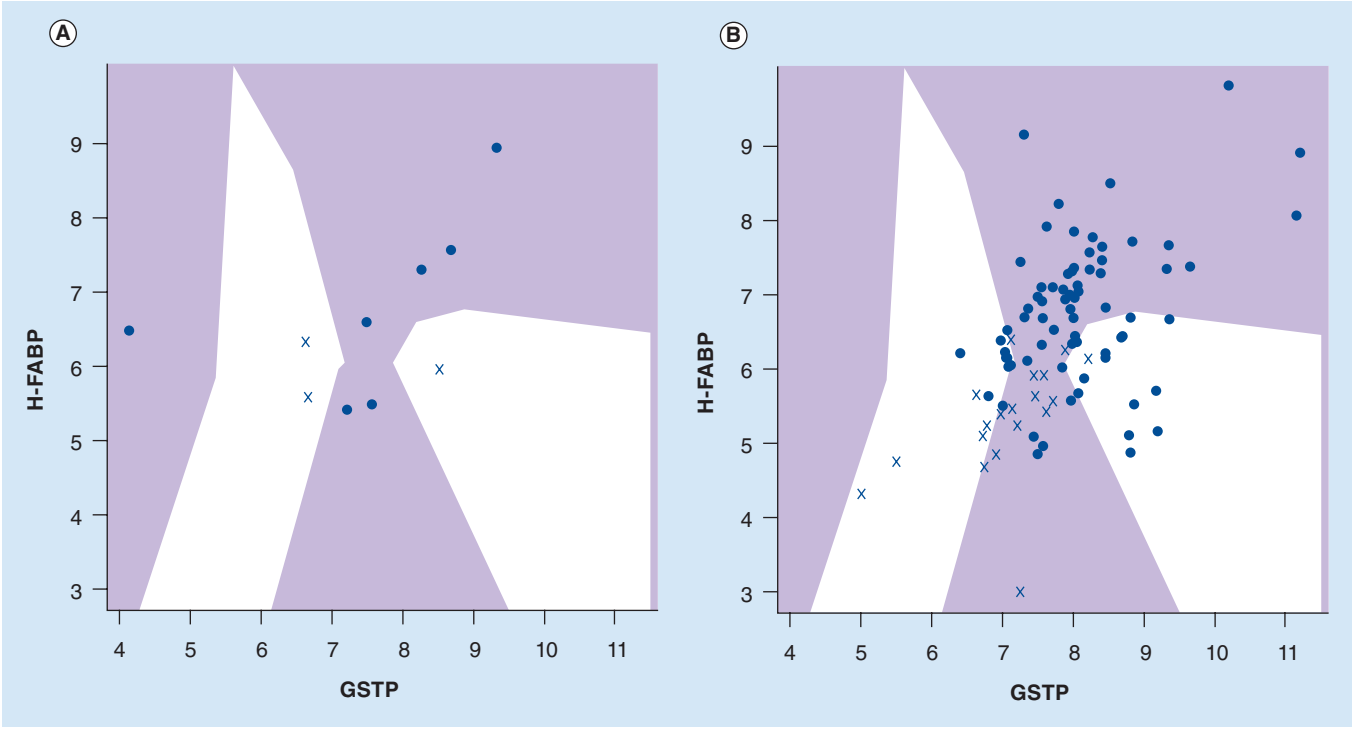


Figure 2. Overfitting with the nearest-neighbour algorithm (k = 1). The gray area shows the region where the test would be considered positive by the method. Crosses and dots represent Stage 1 and Stage 2 human African trypanosomiasis patients, respectively. **(A)** Training set of ten patients determining the class regions (gray or white background). **(B)** The pattern defined in **(A)** is applied to a test set of 90 different patients. Most Stage 1 patients and many Stage 2 patients are misclassified in the test set. The choice of the training and test set is purely for illustrational purposes.
GSTP: Glutathione *S*-transferase Pi; H-FABP: Heart-type fatty acid-binding protein.
Redrawn from [8].

a minimum number. To take a purely theoretical example, one could set a minimum of two out of five parameters, where any two positive molecules of the panel would raise a positive test, but if only one is positive, the panel result would be negative. The minimal number can be chosen based on several criteria, usually depending on the targeted sensitivity or specificity or by cross validation. It is used mostly for ELISA and clinical data but not in higher-dimensional problems. The threshold method has the major advantage that results are easy to interpret. Additionally, its simple boundary structure reduces the possibility of overfitting the training data. In our view, the threshold method is well adapted to biomarker panel data, where class boundaries of a single marker can often be represented as single cut-off points.

Lejon *et al.* followed this approach to combine clinical and biochemical variables to predict trypanosomiasis treatment failure [38]. Thresholds were chosen on univariate parameters to maximize the sum of sensitivity and specificity, and two parameters were retained. For the same disease, Hainard *et al.* selected a panel of two cytokines and a brain-damage marker to assess the disease stage of 100 patients using a multivariate approach [8]. The rationale was that interactions between molecules in a panel can be complex and good univariate thresholds are not necessarily the best thresholds in a panel. Other attempts have been made in this direction [36]. Vitzthum *et al.* also showed that different thresholds should be chosen for different clinical questions [14]. This means that if a threshold discriminates well between classes for one question, it may not automatically be accurate in other problems.

A similar technique is the patient rule-induction method [34], where two thresholds (lower and upper) are chosen, and a patient is positive only if the biomarker value is included in the range. This can bring out patients with particularly low values, but the clinical and biological relevance of such a criterion is not obvious. It was applied by Wang *et al.*, but its usage seems scarce [39]. Naive Bayes is another similar method, in which the thresholds are separately determined based on statistical criteria for every feature. Ralhan *et al.* successfully applied it to proteins quantified by MS/MS after isobaric tag for relative and absolute quantitation labeling on a small number of patients [40]. It can be extended to deal with dependent data [41].

Decision trees

Decision trees are similar to threshold-based methods, but they can find more complex boundaries (Figures 1B & 3 & Box 1). Different tree methods exist and vary in the construction of the tree from the training set, that is, the selection of a feature and a threshold for each node, and in the pruning strategy.

Classification and regression trees (CARTs) are one of the most popular tree-based algorithms [42–44]. Other methods are C4.5 decision trees [45], J48 [46], or recursive partitioning and regression trees (RPART). The latter allowed Ring *et al.* to select five proteins

out of several hundreds and combine them into a decision tree that was able to classify 195 estrogen receptor-positive breast cancer patients into good, moderate or poor prognosis [47]. However, it seemed to be dependent on the cohorts to which the model was applied and was less predictive of outcome than other methods.

Trees perform well in combination with boosting algorithms [48], which can strongly improve the classification results. The idea is to boost the classification performance of a simple classifier (e.g., a strongly pruned tree) by iteratively applying it to modified versions of the data, where the weight of the misclassified training observations is increased. Each successive tree classifier is then forced to focus on those misclassified observations, and the final classification is calculated as the weighted average over all tree classifiers. Trees also form the basis of the random forest algorithm [49], where classification is obtained from a combination of trees, each built from a small but random subset of the features.

A basic parallel or sequential AND/OR way of combining tests similar to decision trees has been proposed by Vitzthum *et al.* [14]. However, there is no evidence that it was applied in panels.

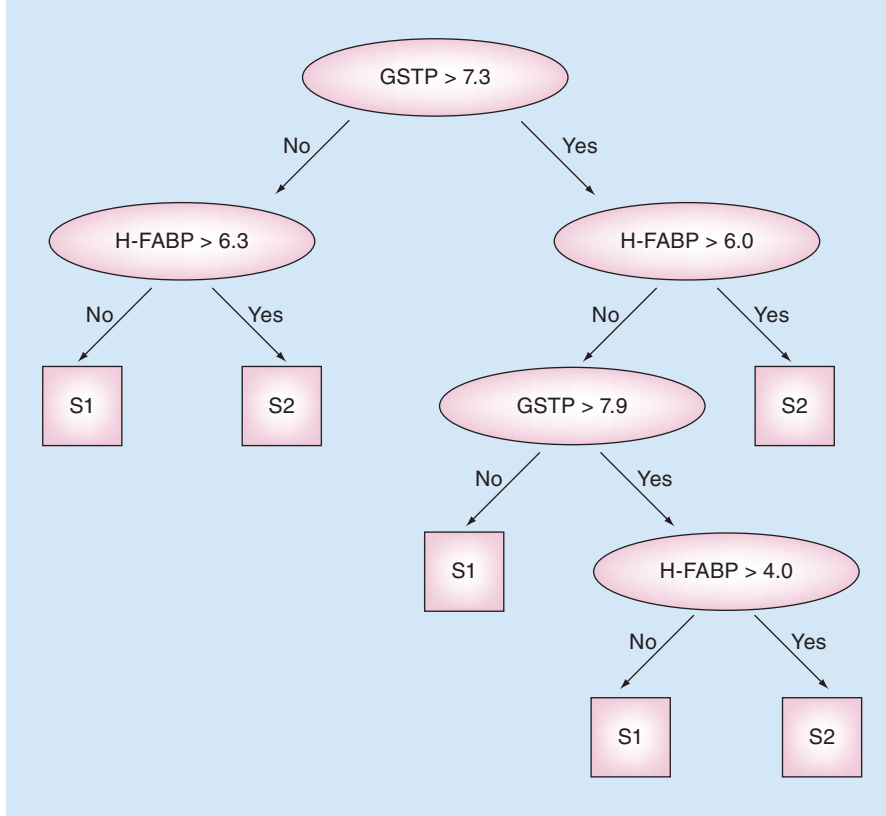


Figure 3. The decision tree corresponding to Figure 1B. Each circle corresponds to a question. Depending on its answer, one follows the arrow to the left or right and goes to the next question, until a decision (square box) is reached.
GSTP: Glutathione *S*-transferase Pi; H-FABP: Heart-type fatty acid-binding protein; S1: Stage 1 human African trypanosomiasis classification; S2: Stage 2 human African trypanosomiasis classification.

Logistic regression

Logistic regression is a very popular linear regression method in the medical field, where the simplicity and robustness of the models produced is appreciated (Figure 1C & Box 2). The method is based on a clear mathematical formulation and yields a globally optimal solution. Interaction terms can be entered to model nonlinear class boundaries, but this requires *a priori* information regarding the structure of the data and, therefore, is not commonly used.

Logistic regression can combine clinical or biomarker data, either continuous or categorical [36,37,45,50–54]. For example, Visintin *et al.* trained several logistic regression models to screen ovarian cancer on several hundred patients and controls [50]. Although some individual biomarkers displayed a significantly lower performance in the test set, regression models were stable, denoting the robustness of the technique. Logistic regression was also applied to combine protein markers with clinical parameters [55] or to combine clinical variables only [56].

Support vector machines

Support vector machines (Figures 1D & 4 & Box 3) are one of the most popular methods in machine learning. SVMs have the advantage of being able to provide a clear mathematical model with a globally optimal solution, contrary to neural networks or others learning methods that can get trapped in a local optimum. It performs well in a large variety of tasks, and it was applied in very different fields, ranging from text pattern recognition to analysis of gene expression microarrays. However the underlying concepts are more difficult to grasp for non-mathematicians. Figure 1D shows the result of classification with a radial basis kernel, but SVMs can also find linear or polynomial separations similar to logistic regression.

The SVM is preferred in higher dimensionality problems, such as microarray [57,58] or mass spectrometry (SELDI [45,46,59] or MALDI [48,60]) data analysis. Liu *et al.* combined use of an SVM with a genetic algorithm and obtained reproducible and fairly accurate results [61]. It was also used by Wild *et al.* to classify ELISA data for patients suffering from rheumatoid arthritis, but only to challenge the regularized discriminant analysis and confirm the results generated by the latter technique [62].

Generalized additive models

Generalized additive models allowed Knickerbocker *et al.* to combine protein microarray data with patient clinical information to predict survival after renal replacement [6]. They added local polynomial functions (or splines) that allow defining nonlinear relationships between the variables, as well as the

Box 1. Decision trees.

- Decision trees are simple but powerful methods that split the feature space into a set of boxes and attribute a class (or a probability) to each one. Figure 3 displays a typical representation of the decision tree corresponding to Figure 1B
- To build a decision tree, a series of binary splits based on a threshold of one of the variables is performed. For each step, the variable that yields the best split is selected. Every outcome of a test (positive or negative) creates a branch, which either leads to a new test or to a terminal leaf, corresponding to a box in the feature space. Each of the boxes is defined by the unique path leading to it, and it is possible to calculate a class probability or binary outcome within the box. The tree is then pruned and the less informative decision branches are removed to simplify the tree and avoid overfitting. The number of splits and the minimal number of observations allowed in each terminal leaf must be carefully investigated, for example, by cross validation [34,90]

detection of inflexion points. They trained two models separately, one for clinical parameters and one for protein biomarkers, and showed that the experimental predictors could only add information for patients detected as being at high risk by the clinical predictors.

Other methods

Several other methods were shown to perform well in proteomics. Gevaert *et al.* applied a Bayesian network on gene expression microarray data [63]. This approach allows the integration of clinical data in several manners: full, decision and partial integration. In full integration, the clinical and microarray datasets are merged and handled as a single dataset. In decision integration, two models are trained, one clinical and one with microarray data, and the final decision is generated as a combination of the weighted probability of the clinical panel with the microarray one. Finally, in partial integration, the network structures are determined separately for each dataset and joined into one single structure before performing the learning step for the merged clinical and microarray datasets.

Regularized discriminant analysis is a classification method that can deal with strongly correlated data [34]. It is based on linear discriminant analysis [48] or quadratic discriminant analysis. It can take into account the main effects of the markers as well as their interaction. Wild *et al.* successfully used regularized discriminant analysis to combine two to three molecules in patients with rheumatoid arthritis [62]. For prognostic purposes, an attractive option is to analyze the time series, if available. James and Hastie proposed a classification based on spline regression of time series and linear discriminant analysis of the regression coefficients [64].

Logical analysis of data is a method that finds approximations of subsets of observations by combinatorics and optimization. Its application in the medical field had been reviewed previously [102]. It was used by Reddy *et al.* to classify 48 ischemic stroke patients and 32 controls, and was applied on a validation set consisting of 60 patients [45]. The methodology was also able to detect two outlier patients and showed good performance.

Reddy *et al.* [45] and Prados *et al.* [46] applied multilayer perceptron, a type of linear neural network, and Cox proportional hazard models. The latter method was used in several other studies [47,52,65,66].

'Nearest neighbors' finds the k nearest samples and performs a majority vote to decide the classification [48]. Linkov *et al.* defined a method they called ADE+PT, which is similar to a weighted nearest-neighbor approach [67]. There is no evidence of its application in any other published study.

Performance validation

Why?

Once a panel is defined, its performance must be evaluated. As stated earlier, overfitting corresponds to the underestimation of the classification error on the training set (Figure 2A), which cannot be validated on an independent test set (Figure 2B) [34]. High-dimensional data are especially prone to overfitting, as mentioned in Feng and Yasui in the context of SELDI mass spectra, where a huge number of possible markers (peptide masses) are available [11]. However, depending on the classifier, it can be a serious problem even for low dimensional data.

Box 2. Logistic regression.

- In its simplest form, logistic regression provides a linear separation of the feature space. It models the class probability p(+|x), that is, the probability that the n-dimensional feature vector, x, is classified positively, as a sigmoidal (s-shaped) function:

$$f(z) = \frac{1}{(1+e^{-z})}$$

where

$$z = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i$$

The coefficients α_i must be determined from the training sample by means of a maximum likelihood procedure, which usually converges to the unique global optimum [34]. If the different features x_i are properly normalized (same mean and standard deviation), the coefficients α_i provide direct information regarding the importance of a feature for the correct classification in the logistic regression model. It is also possible to expand the features by explicitly including interaction and nonlinear terms. For example, the feature vector

$$x = (x_1, x_2)$$

could be expanded to a higher dimensional vector

$$x' = (x_1,\ x_1^2\ +\ x_2^2, x_1x_2, x_2)$$

or

$$x' = (x_1,\ x_1/x_2, x_2)$$

The logistic regression is then applied to x' instead of x

- Odds ratios measure the effect of a given increase of the studied marker. They are frequently used in relation to logistic regression. However, their use as a measure of performance is difficult [91]

Box 3. Support vector machines.

- Let us consider a 2D example where the two classes are completely separable by a straight line. It is easy to see that there are many straight lines that do the job; the question is which of these lines provides the best classification on a test sample? The support vector machine solves this problem by selecting the (usually unique) separating line that is farthest away from any data point [92]. It can be shown that this line often yields better classification performance on a test set since it is as far away as possible from the critical points, which lie close to the class boundary. Mathematically, the linear separation can be formulated as follows: for each feature vector x_i of class y_i (±1) we have we have $w\,x_i + b \leq -1$ for $y_i = -1$ and $w\,x_i + b \geq 1$ for $y_i = 1$ where w is a vector orthogonal to the separating line. It can be shown [92] that the distance of the separating line to the next x_i is $1/|w|$; therefore, the support vector machine searches for the smallest $|w|^2$, which satisfies the aforementioned inequalities. The lines $w\,x_i + b = -1$ for $y_i = -1$ and $w\,x_i + b = 1$ for $y_i = 1$ are termed the margins, which lie parallel and at equal distance $1/|w|$ to the separating line and touch one or more data points of the corresponding class
- In almost all real-life applications, classes are not linearly separable. Cortes and Vapnik, however, showed that a similar approach still works in these cases [92]. They introduced so-called slack variables $\xi_i \geq 0$ and reformulated the constraints as $w\,x_i + b \leq -1+\xi i$ for $y_i = -1$ and $w\,x_i + b \geq 1-\xi i$ for $y_i = 1$, that is, for each x_i on the right side of its margin, we have $\xi_i = 0$, and for each x_i on the wrong side of the margin, $\xi_i > 0$, where $\xi_i/|w|$ is the distance from the margin (Figure 4). Since we still would like to have a margin distance $2/|w|$ as large as possible, but also as little misclassification

$$\sum_{i=1}^{p} \xi_i$$

as possible, we search for a w value that satisfies the 'slack' inequalities mentioned and minimizing

$$|w|^2 + C\sum_{i=1}^{p} \xi_i$$

where p is the number of samples and C a misclassification weight. This is a quadratic programming problem, for which many efficient algorithms are available, which usually converge to a unique solution. It can be shown that

$$w = \sum_{i=1}^{p} \alpha_i y_i x_i$$

where $\alpha_i > 0$ for those sample vectors (so-called support vectors), which lie either on the margin or on the wrong side of it ($w\,x_i + b \geq -1$ for $y_i = -1$ and $w\,x_i + b \leq 1$ for $y_i = 1$), and $\alpha_i = 0$ for all other correctly classified vectors
- Cortes and Vapnik also showed that the support vector machine approach can be naturally extended to nonlinear separation [92]. In Figure 1D for example, we used a radial basis kernel, which yields the class indicator function as a sum over radial basis functions, which are centered at the support vectors (see [34,92] for a detailed discussion of the kernel-based formulation)

In the literature, Bhaskar *et al.* showed that validation is not performed consistently in bioinformatics [68] and Whiteley *et al.* showed how even single biomarkers can be biased if its threshold is chosen on the same dataset [69]. Several panel papers that we previously mentioned did not perform any kind of validation of the accuracy of the reported classification [8,24,35,37,38,52,55] or simply mentioned that it would or should be done later. While this is still acceptable for single biomarkers, doing so with panels could lead to false hopes and should be avoided in the future. Therefore, it is crucial to have a separate dataset that includes patient data, which is independent from the model definition, to test that model. Ideally, the dataset should originate from a separate cohort of patients with biomarker concentration measured in a different laboratory. However, such validation data is often unavailable, and the number of patients is often too small to split the data into independent training and test sets of the same size.

How?

Apart from using an independent validation dataset, which is not always possible, several computational methods can overcome this issue. If the number of patients is sufficient, a subset of the sample population can be left aside for the training process and kept as validation set, which was done by several groups [36,44,50,53]. If not enough patients are available, randomization techniques, such as permutation tests, cross validation and bootstrapping, can help with evaluation if the classification is significant or if it is only overfitting [11].

Permutation tests

Permutation tests allow the determination of whether the classification result is significant [70,71]. Patient labels are randomly permutated, and the problem is treated in the same way, providing information concerning the classification error under the random hypothesis. If the efficiency of the classification of random patients is comparable to that of real patients, it is a strong indication that the method is overfitting the training data.

Cross validation

Cross validation is a purely computational method that allows evaluation of the robustness of a classification. In cross validation, the data are split into a number, *k*, of equal-sized parts. Sequentially, *k*-1 parts are used to train the classifier model, and the remaining one is kept to test the performance of the model. When all parts have been used as test sets, performance is averaged [34].

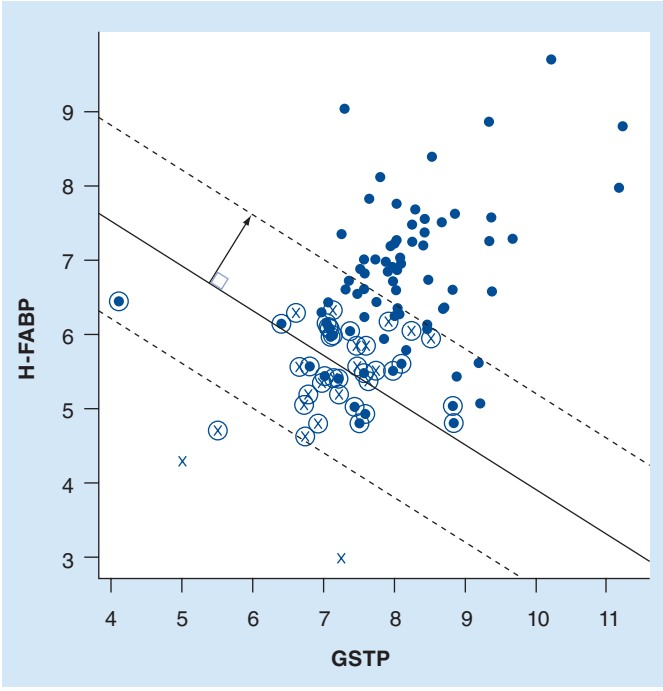


Figure 4. Support vector machines. Crosses and dots represent Stage 1 and Stage 2 human African trypanosomiasis patients, respectively. Margins and the separation line are represented by dashed and solid lines, respectively. Support vector observations are circled in gray. The arrow represents the vector $w/|w|^2$.
GSTP: Glutathione *S*-transferase Pi; H-FABP: Heart-type fatty acid-binding protein.
Redrawn from [8].

Typical values for k are five or ten [34]. If k is equal to the sample size, it is a leave-one-out cross validation. The problem with cross validation is that the training sample size is smaller, which can lead to overestimation of the prediction error. For biologists and clinicians, another problem is that each round of cross validation can choose a different model. Therefore, it must be made clear we evaluate the error of the method, not of the model itself. Several groups applied cross validation for biomarker applications [43,45,47,50,62,67].

Several variants of cross validation exist. When the data are not balanced, that is, one class has a much smaller patient number than the other, a stratified cross validation can be performed, where both classes are represented in the same proportion in each k fold than in the whole set. Another variant is double cross validation, which combines an internal loop where the model meta-parameters (such as the width of a kernel, the kernel-type or the number of principal components) are defined, and an external loop where the model is actually trained with these parameters and performance is evaluated [70].

Bootstrapping

Bootstrapping involves randomly selecting items with replacement in order to obtain a new sample of the same size as the original one. Approximately 37% of the original sample will not be selected and can be used as a test set. This procedure can be repeated a large number of times to get a good approximation [34,72].

In contrast to cross validation, sample size is not reduced but some data will be redundant. It is particularly helpful for determining empirical confidence intervals [73]. Several publications employed bootstrapping for validation [6,42,51]. Similarly to double cross validation, Feng *et al.* proposed that cross validation should serve as model selection and bootstrap as estimation of the classification error [11].

Separate set validation

The ultimate validation is always to reproduce the experiment independently on different patients and within a different laboratory. However, mainly because of time and funding constraints, it cannot always be carried out, and one must rely on previous investigations. For example, Whiteley *et al.* showed that no publication using panels for the diagnosis of ischemic stroke validated its results on an independent patient cohort [70]. They recommended independent validation as a good practice, also for other work dealing with patient classification. Reddy *et al.* [45] and Gevaert *et al.* [63], for example, rely on an independent cohort for validation.

Statistical method reporting

Proteomics is currently moving towards better reporting requirements, such as the 'Minimum Information about a Proteomics Experiment' model [73]. A similar initiative exists in the medical community with the Standards for Reporting of Diagnostic Accuracy that defines a checklist of 25 items to promote a coherent reporting of accuracies [74]. However, none of these initiatives fully covers the needs of panels. As good reporting of panel performance is absolutely required to gain medical community acceptance, we believe that reporting standards will be needed for panels. Detailing what this standard would be is out of the scope of this review, but we can highlight a few points of major importance.

In order to allow the ultimate independent validation by different laboratories, it is very important that statistical analysis methods are discussed in detail and information regarding the software and corresponding parameters is provided. Stating which software was used is important, since default parameters may differ in distinct implementations of the same method. Most studies do not follow this advice, with few exceptions [6,8,37]. For cross validation and bootstrapping, a graph such as that presented by Wild *et al.* usually helps the reader understand how the performance test was applied and what the reported results really mean [62]. Other requirements will need to be discussed by the panel community.

Comparison of methods

As mentioned earlier, several models can be generated from one dataset. Therefore, model comparison is crucial in order to optimize the final selection.

Several papers analyze datasets with more than one method [45,46,48]. However, there is no proper comparison. Reddy *et al.* states that "logical analysis of data model has significantly better performance on the independent validation set compared with the other classification models" [45]. However, there are no statistics to prove this difference, and confidence intervals

partially overlap. Prados *et al.* used McNemar's test for pairwise comparison of algorithms [46]. Wu eludes the problem by studying the stability of the model performance over several cross validation or bootstrap replicates [48].

The most important point is that performance estimates should be compared on a dataset that is independent from the model definition [75,76]. This can be carried out either with an independent validation cohort (separated or split), or by estimating performance via cross validation or bootstrapping.

Receiver-operating characteristic curves

Traditionally, performance of a test discriminating between two classes of patients is evaluated using a receiver-operating characteristic (ROC) curve [77]. This shows the variation of sensitivity and specificity of a test as the decision threshold changes. When the decision threshold is low, sensitivity is high and specificity is low, thus corresponding to the top right zone of the curve. Conversely, when the decision threshold is high, specificity is high and sensitivity is low, which corresponds to the bottom left part of the curve (Figure 5, see also Table 1).

A biomarker with no discrimination power would be characterized by a diagonal line, while a 'perfect' biomarker would reach the top left point corresponding to 100% sensitivity and 100% specificity. A major characteristic of a ROC curve is its AUC. The maximum AUC possible is 100%, corresponding to a 'perfect' classification. A nondiscriminating ROC curve has an AUC of 50%. In 1989, McClish introduced the concept of partial area under the ROC curve [15,78,79]. It consists of analyzing only a region of special interest of the ROC curve and allows the selection of models with high specificity or sensitivity, rather than models with a better average performance but potentially lower clinical value.

Hanley and McNeil [80] and DeLong *et al.* [81] proposed nonparametric methods to compare ROC curves derived from the same sample. McClish described a method to find a specific region within a ROC curve that is different [82]. Baker proposed a method to select best thresholds from a multidimensional ROC curve [83].

An intrinsic property of ROC curves is that the AUC of smooth curves tend to be greater than those of trapezoidal or step graphs [81,84]. Therefore, classification methods or predictors that can take only a few values (such as clinical scores) will not work as well as continuous predictors (such as biomarkers). Several smoothing procedures can be applied to reduce this problem. For example, logistic or other regression techniques will produce smooth estimates of the class probabilities. Gu *et al.* present a smoothing procedure based on Bayesian bootstrap estimation [85].

Another option is to bootstrap and compute confidence intervals and see if the observed sample is compatible with the bootstrap distribution [71,72]. Reddy *et al.* adopted this solution [45].

Classifications

Statistical tests should also be applied in order to judge the significance of differences between classifiers. If only two classifiers are compared, a simple binomial or McNemar test [38,86,87] can calculate the p-value to show that both classifiers are equally good [88]. Both tests are based on a 2 × 2 table, where the diagonal elements count the number of patients where both classifiers agree (either correctly or erroneously), and the off-diagonal elements indicate the number of patients where only one of the classifiers produces the right prediction. The off-diagonal elements are then compared with the calculated p-values. The number of patients where both classifiers agree does not enter into these calculations, which can cause a problem if the number of ties is much larger than the number of discrepancies, and these tests will overestimate the difference between the classifiers. Other, more sophisticated and general tests and methods for testing multiple classifiers are also described in Salzberg's overview [89]. Often, several parameterizations of the same classifiers are tested and the best one is retained. This can lead to overly optimistic results if the p-values are not adjusted for multiple testing. For example, if 20 independent parameterizations are tested at a 5% significance level, one of these parameterizations may exceed the significance level just by chance.

A panel should perform better than each of its individual markers. When comparing the performance of a panel with that of an individual marker, it is important to be as fair as possible. In most publications, the predictions of individual markers are not evaluated by cross validation, which may lead to overly optimistic results [69]. Therefore, we recommend measuring all classifier performances with the same cross validation method or on an independent test set.

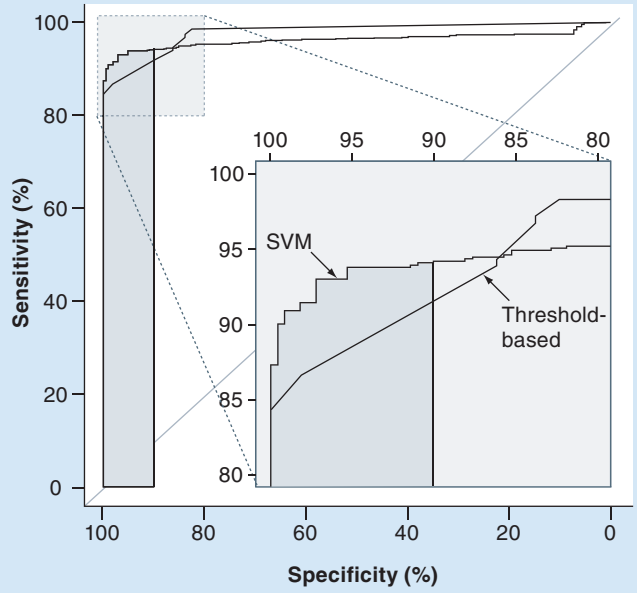


Figure 5. Two receiver-operating characteristic curves for support vector machines and threshold-based classifiers.
Cross validation of partial AUC between 100 and 90% specificity are shown in gray. Respective values of partial AUC are 9.3 and 8.8% (perfect classification would correspond to a partial AUC of 10%).
SVM: Support vector machine.

Expert commentary

Interest in biomarker panels has been growing over the last few years. A number of publications have demonstrated that the approach has a big potential and could be suitable for various clinical applications. They applied many different methods, based on thresholds, decision trees, logistic regression, SVM and several other techniques. None of these methods is clearly superior. SVMs are well studied and tend to work well, even for high-dimensional data, whereas threshold-based methods are easy to implement and understand for medical practitioners. The final choice of a method must be carefully validated.

New markers, although they do not individually perform better than the current ones, could bring useful complementary pieces of information to a panel if they allow evaluation of the state of different pathways. However, such a relation must be sought during the discovery phase, which is made difficult by the very low sample size commonly used.

The limited consensus regarding accepted statistical methods and tools hamper their adoption, and could explain why the number of panels available in clinical practice is still limited. We predict that such standardized methods and tools will soon be made available and that the field will continue to grow despite these current limitations. Validation and comparison are of major importance in the evaluation of panels. It is not always possible to obtain an independent validation cohort, but in this case, the model must be evaluated by cross validation or bootstrap. Here again, the lack of clear guidelines and standards makes it difficult to compare different methods and impedes the credibility of the published results.

Five-year view

To gain a broad acceptance, future panel studies will need to define and follow reporting standards. Special care regarding validation will be required. Robust statistical methods of comparison must still be defined and are a crucial step. There is clearly a critical need for standardized methodologies and reporting standards to gain the medical practitioner's confidence. It is not unreasonable to say that in the absence of a strict enforcement of guidelines, most authors will not comply with better validation and reporting.

In the future, proteomics researchers willing to work with panels will need to think about combinations during the discovery process. Standard feature-selection techniques that select only a few of the best individual markers might reject proteins that are less efficient individually but that might have a greater weight in a panel. Some progress has been made towards this goal, with promising results [89].

We can imagine that proteomics biomarkers, which are still not commonly used in clinical practice, and panels, might contribute to new and more efficient IVD tools. However, given that the field is only in its first stages, it will probably take more than five years to see protein panels used in large scale clinical practice.

Acknowledgements

The authors would like to thank Proteome Sciences Plc for their keen support.

Financial disclosures

Jean-Charles Sanchez receives a grant from Proteome Sciences Plc for brain marker discovery. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Key issues

- A panel is the combination of information from several molecules into one predictor.
- Several methods can be applied. None of them is clearly superior. Support vector machines are usually preferred for high-dimensional data, such as mass spectra, while logistic regression or threshold-based methods are commonly preferred with ELISA-measured biomarkers.
- Methods are difficult to compare, and no efficient comparison tool is available yet.
- An especially careful validation is required in order not to overestimate the performance. It can be achieved either by using a separate dataset or by means of cross validation and/or bootstrap. Validation in an independent cohort measured by a different group is eventually required.
- Reporting detailed information regarding software and parameters set for preprocessing, classification, validation and comparison of methods should be seen as requirements. Reporting standards need to be developed.

References

Papers of special note have been highlighted as:
- • of interest
- •• of considerable interest

1 Hanash S. Disease proteomics. *Nature* 422(6928), 226–232 (2003).

2 Steel LF, Haab BB, Hanash SM. Methods of comparative proteomic profiling for disease diagnostics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 815(1–2), 275–284 (2005).

3 Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science* 312(5771), 212–217 (2006).

4 Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. *Nature* 452(7187), 571–579 (2008).

5 Panchaud A, Affolter M, Moreillon P, Kussmann M. Experimental and computational approaches to quantitative proteomics: status quo and outlook. *J. Proteomics* 71(1), 19–33 (2008).

6 Knickerbocker T, Chen JR, Thadhani R, MacBeath G. An integrated approach to prognosis using protein microarrays and nonparametric methods. *Mol. Syst. Biol.* 3(123), 1–8 (2007).

- • **A complete study with a multiplex protein array approach and a thorough statistical analysis.**

7 Turck N, Vutskits L, Sanchez-Pena P *et al.* A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Med.* DOI: 10.1007/s00134-009-1641-y (2009) (Epub ahead of print).

8 Hainard A, Tiberti N, Robin X *et al.* A combined CXCL10, CXCL8 and H-FABP panel for the staging of human African trypanosomiasis patients. *PLoS Negl. Trop. Dis.* 3(6), E459 (2009).

- **Impressive results on a neglected tropical disease and a potential clinical application in the short to medium term.**

9 Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 20(5), 777–785 (2004).

10 Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J. Natl Cancer Inst.* 96(5), 353–356 (2004).

11 Feng Z, Yasui Y. Statistical considerations in combining biomarkers. *Dis. Markers* 20(2), 45–51 (2004).

12 Stead DA, Paton NW, Missier P *et al.* Information quality in proteomics. *Brief Bioinform.* 9(2), 174–188 (2008).

13 Oberg AL, Vitek O. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J. Proteome Res.* 8(5), 2144–2156 (2009).

14 Vitzthum F, Behrens F, Anderson NL, Shaw JH. Proteomics: from basic research to diagnostic application. A review of requirements and needs. *J. Proteome Res.* 4(4), 1086–1097 (2005).

15 Dodd LE, Pepe MS. Partial AUC estimation and regression. *Biometrics* 59(3), 614–623 (2003).

16 Hackett JL, Gutman SI. Introduction to the Food and Drug Administration (FDA) regulatory process. *J. Proteome Res.* 4(4), 1110–1113 (2005).

17 Vanni S, Polidori G, Pepe G *et al.* Use of biomarkers in triage of patients with suspected stroke. *J. Emerg. Med.* DOI: 10.1016/j.jemermed.2008.09.028 (2009) (Epub ahead of print).

18 Sibon I, Rouanet F, Meissner W, Orgogozo JM. Use of the Triage Stroke Panel in a neurologic emergency service. *Am. J. Emerg. Med.* 27(5), 558–562 (2009).

19 Brouns R, Van Den Bossche J, De Surgeloose D, Sheorajpanday R, De Deyn PP. Clinical and biochemical diagnosis of small-vessel disease in acute ischemic stroke. *J. Neurol. Sci.* 285(1–2), 185–190 (2009).

20 Apple FS, Christenson RH, Valdes R *et al.* Simultaneous rapid measurement of whole blood myoglobin, creatine kinase MB, and cardiac troponin I by the triage cardiac panel for detection of myocardial infarction. *Clin. Chem.* 45(2), 199–205 (1999).

21 Hoffer A, Osmond H. A card sorting test helpful in making psychiatric diagnosis. *J. Neuropsychiatr.* 2, 306–330 (1961).

22 Report of World Federation of Neurological Surgeons Committee on a Universal Subarachnoid Hemorrhage Grading Scale. *J. Neurosurg.* 68(6), 985–986 (1988).

23 Woolas RP, Xu F-J, Jacabs IJ *et al.* Elevation of multiple serum markers in patients with stage I ovarian cancer. *J. Natl Cancer Inst.* 85(21), 1748–1751 (1993).

24 Hill MD, Jackowski G, Bayer N, Lawrence M, Jaeschke R. Biochemical markers in acute ischemic stroke. *Can. Med. Assoc. J.* 162(8), 1139–1140 (2000).

25 Ferguson RE, Hochstrasser DF, Banks RE. Impact of preanalytical variables on the analysis of biological fluids in proteomic studies. *Proteomics Clin. Appl.* 1(8), 739–746 (2007).

26 Rai AJ, Vitzthum F. Effects of preanalytical variables on peptide and protein measurements in human serum and plasma: implications for clinical proteomics. *Expert Rev. Proteomics* 3(4), 409–426 (2006).

27 Little RR, Rohlfing CL, Tennill AL *et al.* Standardization of C-peptide measurements. *Clin. Chem.* 54(6), 1023–1026 (2008).

28 Yeo I-K, Johnson RA. A new family of power transformations to improve normality or symmetry. *Biometrika* 87(4), 954–959 (2000).

29 Gentleman R, Huber W, Carey V, Irizarry RA, Dudoit S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Gentleman R, Huber W, Carey V, Irizarry RA, Dudoit S (Eds). Springer-Verlag, NY, USA (2005).

30 Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. *Brief Bioinform.* 9(2), 102–118 (2008).

31 Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19), 2507–2517 (2007).

32 Baggerly KA, Morris JS, Wang J, Gold D, Xiao L-C, Coombes KR. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* 3(9), 1667–1672 (2003).

33 Petricoin EF, Ardekani AM, Hitt BA *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359(9306), 572–577 (2002).

34 Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning: data mining, inference, and prediction.* Springer-Verlag, NY, USA (2003).

- **Reference book for statistical learning.**

35 Faca VM, Song KS, Wang H *et al.* A mouse to human search for plasma proteome changes associated with pancreatic tumor development. *PLoS Med.* 5(6), E123 (2008).

36 Reynolds MA, Kirchick HJ, Dahlen JR *et al.* Early biomarkers of stroke. *Clin. Chem.* 49(10), 1733–1739 (2003).

- **First attempt to select a panel with thresholds defined in a multivariate manner.**

37 Montaner J, Perea-Gainza M, Delgado P *et al.* Etiologic diagnosis of ischemic stroke subtypes with plasma biomarkers. *Stroke* 39(8), 2280–2287 (2008).

38 Lejon V, Roger I, Mumba Ngoyi D *et al.* Novel markers for treatment outcome in late-stage *Trypanosoma brucei gambiense* trypanosomiasis. *Clin. Infect. Dis.* 47(1), 15–22 (2008).

39 Wang P, Kim Y, Pollack J, Tibshirani R. Boosted PRIM with application to searching for oncogenic pathway of lung cancer. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference.* Stanford, CA, USA, 16–19 August 2004.

40 Ralhan R, DeSouza LV, Matta A *et al.* Discovery and verification of head-and-neck cancer biomarkers by differential protein expression analysis using iTRAQ labeling, multidimensional liquid chromatography, and tandem mass spectrometry. *Mol. Cell. Proteomics* 7(6), 1162–1173 (2008).

41 Webb GI, Boughton JR, Wang Z. Not so naive Bayes: aggregating one-dependence estimators. *Machine Learn.* 58(1), 5–24 (2005).

42 Seeber B, Sammel MD, Fan X *et al.* Panel of markers can accurately predict endometriosis in a subset of patients. *Fertil. Steril.* 89(5), 1073–1081 (2008).

43 Rosen RC, Cappelleri JC, Smith MD, Lipsky J, Peña BM. Development and evaluation of an abridged 5-item version of the International Index of Erectile Function (IIEF-5) as a diagnostic tool for erectile dysfunction. *Int. J. Impot. Res.* 11(6), 319–326 (1999).

44 Patz EF, Campa MJ, Gottlin EB, Kusmartseva I, Guan XR, Herndon JE. Panel of serum biomarkers for the diagnosis of lung cancer. *J. Clin. Oncol.* 25(35), 5578–5583 (2007).

45 Reddy A, Wang H, Yu H *et al.* Logical Analysis of Data (LAD) model for the early diagnosis of acute ischemic stroke. *BMC Med. Inform. Decis. Mak.* 8, 30 (2008).

•• **Comparison of five methods (Logical Analysis of Data, support vector machine, decision tree, logistic regression and multilayer perceptron) on SELDI data. Exemplary validation with tenfold cross-validation and an independent validation set.**

46 Prados J, Kalousis A, Sanchez J-C, Allard L, Carrette O, Hilario M. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics* 4(8), 2320–2332 (2004).

47 Ring BZ, Seitz RS, Beck R *et al.* Novel prognostic immunohistochemical biomarker panel for estrogen receptor-positive breast cancer. *J. Clin. Oncol.* 24(19), 3039–3047 (2006).

48 Wu B, Abbott T, Fishman D *et al.* Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19(13), 1636–1643 (2003).

49 Breiman L. Random forests. *Machine Learn.* 45(1), 5–32 (2001).

50 Visintin I, Feng Z, Longton G *et al.* Diagnostic markers for early detection of ovarian cancer. *Clin. Cancer Res.* 14(4), 1065–1072 (2008).

•• **Describes several logistic regression models fitted with a very clear validation. Data were acquired in a multiplex bead assay.**

51 Lynch JR, Blessing R, White WD, Grocott HP, Newman MF, Laskowitz DT. Novel diagnostic test for acute stroke. *Stroke* 35(1), 57–63 (2004).

52 Rosengart AJ, Schultheiss KE, Tolentino J, Macdonald RL. Prognostic factors for outcome in patients with aneurysmal subarachnoid hemorrhage. *Stroke* 38(8), 2315–2321 (2007).

53 Zheng Y, Katsaros D, Shan SJC *et al.* A multiparametric panel for ovarian cancer diagnosis, prognosis, and response to chemotherapy. *Clin. Cancer Res.* 13(23), 6984–6992 (2007).

54 Laskowitz DT, Kasner SE, Saver J, Remmel KS, Jauch EC; BRAIN Study Group. Clinical usefulness of a biomarker-based diagnostic test for acute stroke: the Biomarker Rapid Assessment in Ischemic Injury (BRAIN) study. *Stroke* 40(1), 77–85 (2009).

55 Welsh P, Barber M, Langhorne P, Rumley A, Lowe GDO, Stott DJ. Associations of inflammatory and haemostatic biomarkers with poor outcome in acute ischaemic stroke. *Cerebrovasc. Dis.* 27(3), 247–253 (2009).

56 Wicki J, Perneger TV, Junod AF, Bounameaux H, Perrier A. Assessing clinical probability of pulmonary embolism in the emergency ward: a simple score. *Arch. Intern. Med.* 161(1), 92–97 (2001).

57 Zervakis M, Blazadonakis ME, Tsiliki G, Danilatou V, Tsiknakis M, Kafetzopoulos D. Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC Bioinformatics* 10, 53 (2009).

58 Schramm A, Schulte JH, Klein-Hitpass L *et al.* Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene* 24(53), 7902–7912 (2005).

59 Petricoin EF, Liotta LA. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. *Curr. Opin. Biotechnol.* 15(1), 24–30 (2004).

60 Ressom HW, Varghese RS, Drake SK *et al.* Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* 23(5), 619–626 (2007).

61 Liu JJ, Cutler G, Li W *et al.* Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics* 21(11), 2691–2697 (2005).

62 Wild N, Karl J, Grunert VP *et al.* Diagnosis of rheumatoid arthritis: multivariate analysis of biomarkers. *Biomarkers* 13(1), 88–105 (2008).

63 Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22(14), E184–E190 (2006).

64 James GM, Hastie TJ. functional linear discriminant analysis for irregularly sampled curves. *J. R. Statist. Soc. B* 63(3), 533–550 (2001).

65 Ishino M, Takeishi Y, Niizeki T *et al.* Implications of BNP, H-FABP, and PTX3. *Circ. J.* 72(11), 1800–1805 (2008).

66 Ross DT, Kim C-Y, Tang G *et al.* Chemosensitivity and stratification by a five monoclonal antibody immunohistochemistry test in the NSABP B14 and B20 trials. *Clin. Cancer Res.* 14(20), 6602–6609 (2008).

67 Linkov F, Lisovich A, Yurkovetsky Z *et al.* Early detection of head and neck cancer: development of a novel screening tool using multiplexed immunobead-based biomarker profiling. *Cancer Epidemiol. Biomarkers Prev.* 16(1), 102–107 (2007).

68 Bhaskar H, Hoyle DC, Singh S. Machine learning in bioinformatics: a brief survey and recommendations for practitioners. *Comput. Biol. Med.* 36(10), 1104–1125 (2006).

69 Whiteley W, Tseng M-C, Sandercock P. Blood biomarkers in the diagnosis of ischemic stroke: a systematic review. *Stroke* 39(10), 2902–2909 (2008).

70 Smit S, van Breemen MJ, Hoefsloot HCJ, Smilde AK, Aerts JMFG, de Koster CG. Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta* 592(2), 210–217 (2007).

71 Hesterberg T, Moore DS, Monaghan S, Clipson A, Epstein R. Bootstrap methods and permutation tests. In: *Introduction to the Practice of Statistics (5th Edition).* WH Freeman & Company, NY, USA (2005).

72 Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.* 19(9), 1141–1164 (2000).

73 Taylor CF. Minimum reporting requirements for proteomics: a MIAPE primer. *Proteomics* 6(Suppl. 2), 39–44 (2006).

74 Bossuyt PM, Reitsma JB, Bruns DE *et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin. Chem.* 49(1), 7–18 (2003).

75 LaBaer J. So, you want to look for biomarkers (introduction to the special biomarkers issue). *J. Proteome Res.* 4(4), 1053–1059 (2005).

76 Hilario M, Kalousis A, Pellegrini C, Müller M. Processing and classification of protein mass spectra. *Mass Spectrom. Rev.* 25(3), 409–449 (2006).

77 Fawcett T. An introduction to ROC analysis. *Pattern Recogn. Lett.* 27(8), 861–874 (2006).

78 McClish DK. Analyzing a portion of the ROC Curve. *Med. Decis. Making* 9(3), 190–195 (1989).

79 Thompson ML, Zucchini W. On the statistical analysis of ROC curves. *Stat. Med.* 8(10), 1277–1290 (1989).

80 Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148(3), 839–843 (1983).

81 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3), 837–845 (1988).

82 McClish DK. Determining a range of false-positive rates for which ROC curves differ. *Med. Decis. Making* 10(4), 283–287 (1990).

83 Baker SG. Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* 56(4), 1082–1087 (2000).

84 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1), 29–36 (1982).

85 Gu J, Ghosal S, Roy A. Bayesian bootstrap estimation of ROC curve. *Stat. Med.* 27(26), 5407–5420 (2008).

86 Morais DF, Spotti AR, Tognola WA, Gaia FFP, Andrade AF. Clinical application of magnetic resonance in acute traumatic brain injury. *Arquivos de Neuro-Psiquiatria* 66(1), 53–58 (2008).

87 Vasconcelos OM Jr, Prokhorenko OA, Kelley KF *et al.* A comparison of fatigue scales in postpoliomyelitis syndrome. *Arch. Phys. Med. Rehabil.* 87(9), 1213–1217 (2006).

88 Salzberg S. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Min. Knowl. Discov.* 1(3), 317–328 (1997).

89 Gillette MA, Mani DR, Carr SA. Place of pattern in proteomic biomarker discovery. *J. Proteome Res.* 4(4), 1143–1154 (2005).

90 Han J, Kamber M. *Data Mining.* Morgan Kaufmann, CA, USA (2001).

91 Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am. J. Epidemiol.* 159(9), 882–890 (2004).

92 Cortes C, Vapnik V. Support-vector networks. *Machine Learn.* 20(3), 273–297 (1995).

Websites

101 McCormick T, Martin K, Hehenberger M. The evolving role of biomarkers – Focusing on patients from research to clinical practice. *IBM Global Business Services* (2007) www.03.ibm.com/industries/healthcare/doc/content/resource/insight/2799019105.html

102 Hammer PL, Bonates T. Logical analysis of data: from combinatorial optimization to medical applications. RUTCOR Research Report (10–2005) (2005) http://rutcor.rutgers.edu/pub/rrr/reports2005/10_2005.pdf

Affiliations

- Xavier Robin
  Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
- Natacha Turck
  Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
- Alexandre Hainard
  Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
- Frédérique Lisacek
  Swiss Institute of Bioinformatics, Medical University Centre, Geneva, Switzerland
- Jean-Charles Sanchez
  Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
  jean-charles.sanchez@unige.ch
- Markus Müller
  Swiss Institute of Bioinformatics, Medical University Centre, Geneva, Switzerland

3

pROC: an open-source package for R and S+ to analyze and compare ROC curves

As outlined in the previous chapter, receiver operating characteristic (ROC) curves are commonly used in biomedical and bioinformatics applications to evaluate the performance of classifiers. Unfortunately, we found that the statistical analysis is often insufficient to support the claims made in the literature. Therefore, to be able to perform valid ROC analysis, we developed a package for R and S+ , called pROC, and released it under an open-source license.

pROC is a set of tools that enables the analysis, display, smoothing and comparison of ROC curves. The R version of the package is written with user-friendly, object-oriented interfaces. The S+ version additionally features a graphical user interface (GUI) for users with no programming skills. With data imported into the R or S+ environments in a standard manner, pROC computes the characteristics (sensitivity and specificity) of the ROC curve. It features several analysis functions for the computation full or partial area under the ROC curve (AUC), confidence intervals, and methods to statistically compare two ROC curves. Finally, intermediary results can be queried and visualized in user-friendly interfaces.

We show a case-study to demonstrate how a ROC analysis can be performed with pROC, based on the data first published in chapter 5.

In conclusion, pROC is a package for R and S+ specifically dedicated to ROC analysis. It features several statistical tests to compare ROC curves, and in particular partial areas under the curve, allowing proper ROC interpretation. pROC is available in two versions: in the R programming language or with a graphical user interface in the S+ statistical software. It is accessible at expasy.org/tools/pROC/ under the GNU General Public License. It is also distributed through the CRAN and CSAN public repositories, facilitating its installation.

This article is an important part of my thesis. I carried out the programming, packaged the code, analyzed the data and fully wrote the manuscript, with invaluable advice from the co-authors.

BMC Bioinformatics

SOFTWARE Open Access

pROC: an open-source package for R and S+ to analyze and compare ROC curves

Xavier Robin[1*], Natacha Turck[1], Alexandre Hainard[1], Natalia Tiberti[1], Frédérique Lisacek[2], Jean-Charles Sanchez[1] and Markus Müller[2*]

Abstract

Background: Receiver operating characteristic (ROC) curves are useful tools to evaluate classifiers in biomedical and bioinformatics applications. However, conclusions are often reached through inconsistent use or insufficient statistical analysis. To support researchers in their ROC curves analysis we developed *pROC*, a package for R and S+ that contains a set of tools displaying, analyzing, smoothing and comparing ROC curves in a user-friendly, object-oriented and flexible interface.

Results: With data previously imported into the R or S+ environment, the *pROC* package builds ROC curves and includes functions for computing confidence intervals, statistical tests for comparing total or partial area under the curve or the operating points of different classifiers, and methods for smoothing ROC curves. Intermediary and final results are visualised in user-friendly interfaces. A case study based on published clinical and biomarker data shows how to perform a typical ROC analysis with *pROC*.

Conclusions: *pROC* is a package for R and S+ specifically dedicated to ROC analysis. It proposes multiple statistical tests to compare ROC curves, and in particular partial areas under the curve, allowing proper ROC interpretation. *pROC* is available in two versions: in the R programming language or with a graphical user interface in the S+ statistical software. It is accessible at http://expasy.org/tools/pROC/ under the GNU General Public License. It is also distributed through the CRAN and CSAN public repositories, facilitating its installation.

Background

A ROC plot displays the performance of a binary classification method with continuous or discrete ordinal output. It shows the sensitivity (the proportion of correctly classified positive observations) and specificity (the proportion of correctly classified negative observations) as the output threshold is moved over the range of all possible values. ROC curves do not depend on class probabilities, facilitating their interpretation and comparison across different data sets. Originally invented for the detection of radar signals, they were soon applied to psychology [1] and medical fields such as radiology [2]. They are now commonly used in medical decision making, bioinformatics [3], data mining and machine learning, evaluating biomarker performances or comparing scoring methods [2,4].

In the ROC context, the area under the curve (AUC) measures the performance of a classifier and is frequently applied for method comparison. A higher AUC means a better classification. However, comparison between AUCs is often performed without a proper statistical analysis partially due to the lack of relevant, accessible and easy-to-use tools providing such tests. Small differences in AUCs can be significant if ROC curves are strongly correlated, and without statistical testing two AUCs can be incorrectly labelled as similar. In contrast a larger difference can be non significant in small samples, as shown by Hanczar *et al.* [5], who also provide an analytical expression for the variance of AUC's as a function of the sample size. We recently identified this lack of proper statistical comparison as a potential cause for the poor acceptance of biomarkers as diagnostic tools in medical applications [6]. Evaluating a classifier by means of total AUC is not suitable when

* Correspondence: Xavier.Robin@unige.ch; markus.mueller@isb-sib.ch
[1]Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
[2]Swiss Institute of Bioinformatics, Medical University Centre, Geneva, Switzerland
Full list of author information is available at the end of the article

the performance assessment only takes place in high specificity or high sensitivity regions [6]. To account for these cases, the partial AUC (pAUC) was introduced as a local comparative approach that focuses only on a portion of the ROC curve [7-9].

Software for ROC analysis already exists. A previous review [10] compared eight ROC programs and found that there is a need for a tool performing valid and standardized statistical tests with good data import and plot functions.

The R [11] and S+ (TIBCO Spotfire S+ 8.2, 2010, Palo Alto, CA) statistical environments provide an extensible framework upon which software can be built. No ROC tool is implemented in S+ yet while four R packages computing ROC curves are available:

1) *ROCR* [12] provides tools computing the performance of predictions by means of precision/recall plots, lift charts, cost curves as well as ROC plots and AUCs. Confidence intervals (CI) are supported for ROC analysis but the user must supply the bootstrapped curves.

2) The *verification* package [13] is not specifically aimed at ROC analysis; nonetheless it can plot ROC curves, compute the AUC and smooth a ROC curve with the binomial model. A Wilcoxon test for a single ROC curve is also implemented, but no test comparing two ROC curves is included.

3) Bioconductor includes the *ROC* package [14] which can only compute the AUC and plot the ROC curve.

4) Pcvsuite [15] is an advanced package for ROC curves which features advanced functions such as covariate adjustment and ROC regression. It was originally designed for Stata and ported to R. It is not available on the CRAN (comprehensive R archive network), but can be downloaded for Windows and MacOS from http://labs.fhcrc.org/pepe/dabs/rocbasic.html.

Table 1 summarizes the differences between these packages. Only pcvsuite enables the statistical comparison between two ROC curves. Pcvsuite, ROCR and ROC can compute AUC or pAUC, but the pAUC can only be defined as a portion of specificity.

The *pROC* package was designed in order to facilitate ROC curve analysis and apply proper statistical tests for their comparison. It provides a consistent and user-friendly set of functions building and plotting a ROC curve, several methods smoothing the curve, computing the full or partial AUC over any range of specificity or sensitivity, as well as computing and visualizing various CIs. It includes tests for the statistical comparison of two ROC curves as well as their AUCs and pAUCs. The software comes with an extensive documentation and relies on the underlying R and S+ systems for data input and plots. Finally, a graphical user interface (GUI) was developed for S+ for users unfamiliar with programming.

Implementation

AUC and pAUC

In *pROC*, the ROC curves are empirical curves in the sensitivity and specificity space. AUCs are computed with trapezoids [4]. The method is extended for pAUCs by ignoring trapezoids outside the partial range and adding partial trapezoids with linear interpolation when necessary. The pAUC region can be defined either as a portion of specificity, as originally described by McClish [7], or as a portion of sensitivity, as proposed later by Jiang *et al.* [8]. Any section of the curve pAUC(t_0, t_1) can be analyzed, and not only portions anchored at 100% specificity or 100% sensitivity. Optionally, pAUC can be standardized with the formula by McClish [7]:

$$\frac{1}{2}\left(1+\frac{pAUC-\min}{\max-\min}\right), \quad (1)$$

where *min* is the pAUC over the same region of the diagonal ROC curve, and *max* is the pAUC over the

Table 1 Features of the R packages for ROC anaylsis

Package name	ROCR	Verification	ROC (Bioconductor)	pcvsuite	pROC
Smoothing	No	Yes	No	Yes	Yes
Partial AUC	Only SP[1]	No	Only SP[1]	Only SP	SP and SE
Confidence intervals	Partial[2]	Partial[3]	No	Partial[4]	Yes
Plotting Confidence Intervals	Yes	Yes	No	Yes	Yes
Statistical tests	No	AUC (one sample)	No	AUC, pAUC, SP	AUC, pAUC, SP, SE, ROC
Available on CRAN	Yes	Yes	No, http://www.bioconductor.org/	No, http://labs.fhcrc.org/pepe/dabs/	Yes

[1]Partial AUC only between 100% and a specified cutoff of specificity.

[2]Bootstrapped ROC curves must be computed by the user.

[3]Only threshold averaging.

[4]Only at a given specificity or inverse ROC.

same region of the perfect ROC curve. The result is a standardized pAUC which is always 1 for a perfect ROC curve and 0.5 for a non-discriminant ROC curve, whatever the partial region defined.

Comparison

Two ROC curves are "paired" (or sometimes termed "correlated" in the literature) if they derive from multiple measurements on the same sample. Several tests exist to compare paired [16-22] or unpaired [23] ROC curves. The comparison can be based on AUC [16-19,21], ROC shape [20,22,23], a given specificity [15] or confidence bands [3,24]. Several tests are implemented in *pROC*. Three of them are implemented without modification from the literature [17,20,23], and the others are based on the bootstrap percentile method.

The bootstrap test to compare AUC or pAUC in *pROC* implements the method originally described by Hanley and McNeil [16]. They define Z as

$$Z = \frac{\theta_1 - \theta_2}{sd\,(\theta_1 - \theta_2)}, \tag{2}$$

where θ_1 and θ_2 are the two (partial) AUCs. Unlike Hanley and McNeil, we compute $sd(\theta_1 - \theta_2)$ with N (defaults to 2000) bootstrap replicates. In each replicate r, the original measurements are resampled with replacement; both new ROC curves corresponding to this new sample are built, the resampled AUCs $\theta_{1,r}$ and $\theta_{2,r}$ and their difference $D_r = \theta_{1,r} - \theta_{2,r}$ are computed. Finally, we compute $sd(\theta_1 - \theta_2) = sd(D)$. As Z approximately follows a normal distribution, one or two-tailed p-values are calculated accordingly. This bootstrap test is very flexible and can be applied to AUC, pAUC and smoothed ROC curves.

Bootstrap is stratified by default; in this case the same number of case and control observations than in the original sample will be selected in each bootstrap replicate. Stratification can be disabled and observations will be resampled regardless of their class labels. Repeats for the bootstrap and progress bars are handled by the *plyr* package [25].

The second method to compare AUCs implemented in *pROC* was developed by DeLong et al. [17] based on U-statistics theory and asymptotic normality. As this test does not require bootstrapping, it runs significantly faster, but it cannot handle pAUC or smoothed ROC curves. For both tests, since the variance depends on the covariance of the ROC curves (Equation 3), strongly correlated ROC curves can have similar AUC values and still be significantly different.

$$\mathrm{var}\,(\theta_1 - \theta_2) = \mathrm{var}\,(\theta_1) + \mathrm{var}\,(\theta_2) - 2\,\mathrm{cov}\,(\theta_1, \theta_2) \tag{3}$$

Venkatraman and Begg [20] and Venkatraman [23] introduced tests to compare two actual ROC curves as opposed to their respective AUCs. Their method evaluates the integrated absolute difference between the two ROC curves, and a permutation distribution is generated to compute the statistical significance of this difference. As the measurements leading to the two ROC curves may be performed on different scales, they are not generally exchangeable between two samples. Therefore, the permutations are based on ranks, and ranks are recomputed as described in [20] to break the ties generated by the permutation.

Finally a test based on bootstrap is implemented to compare the ROC curve at a given level of specificity or sensitivity as proposed by Pepe *et al.* [15]. It works similar to the (p)AUC test, but instead of computing the (p)AUC at each iteration, the sensitivity (or specificity) corresponding to the given specificity (or respectively sensitivity) is computed. This test is equivalent to a pAUC test with a very small pAUC range.

Confidence intervals

CIs are computed with Delong's method [17] for AUCs and with bootstrap for pAUCs [26]. The CIs of the thresholds or the sensitivity and specificity values are computed with bootstrap resampling and the averaging methods described by Fawcett [4]. In all bootstrap CIs, patients are resampled and the modified curve is built before the statistics of interest is computed. As in the bootstrap comparison test, the resampling is done in a stratified manner by default.

Smoothing

Several methods to smooth a ROC curve are also implemented. Binormal smoothing relies on the assumption that there exists a monotone transformation to make both case and control values normally distributed [2]. Under this condition a simple linear relationship (Equation 4) holds between the normal quantile function (ϕ) values of sensitivities and specificities. In our implementation, a linear regression between all quantile values defines a and b, which then define the smoothed curve.

$$\phi^{-1}(SE) = a + b\phi^{-1}(SP) \tag{4}$$

This is different from the method described by Metz et al. [27] who use maximum likelihood estimation of a and b. Binormal smoothing was previously shown to be robust and to provide good fits in many situations even when the deviation from basic assumptions is quite strong [28]. For continuous data we also include methods for kernel (density) smoothing [29], or to fit various known distributions to the class densities with *fitdistr* in the MASS package [30]. If a user would like to run a custom smoothing algorithm that is optimized for the

analysed data, then *pROC* also accepts class densities or the customized smoothing function as input. CI and statistical tests of smoothed AUCs are done with bootstrap.

Results and Discussion

We first evaluate the accuracy of the ROC comparison tests. Results in Additional File 1 show that all unpaired tests give uniform p-values under a null hypothesis (Additional Files 1 and 2) and that there is a very good correlation between DeLong's and bootstrap tests (Additional Files 1 and 3). The relation between Venkatraman's and the other tests is also investigated (Additional Files 1 and 4).

We now present how to perform a typical ROC analysis with *pROC*. In a recent study [31], we analyzed the level of several biomarkers in the blood of patients at hospital admission after aneurysmal subarachnoid haemorrhage (aSAH) to predict the 6-month outcome. The 141 patients collected were classified according to their outcome with a standard neurological scale, the Glasgow outcome scale (GOS). The biomarker performances were compared with the well established neurological scale of the World Federation of Neurological Surgeons (WFNS), also obtained at admission.

Case study on clinical aSAH data

The purpose of the case presented here is to identify patients at risk of poor post-aSAH outcome, as they require specific healthcare management; therefore the clinical test must be highly specific. Detailed results of the study are reported in [31]. We only outline the features relevant to the ROC analysis.

ROC curves were generated in *pROC* for five biomarkers (H-FABP, S100β, Troponin I, NKDA and UFD-1) and three clinical factors (WFNS, Modified Fisher score and age).

AUC and pAUC

Since we are interested in a clinical test with a high specificity, we focused on partial AUC between 90% and 100% specificity.

The best pAUC is obtained by WFNS, with 3.1%, closely followed by S100β with 3.0% (Figure 1). A perfect clinical test within the same region corresponds to a pAUC of 10%, while a ROC curve without any discrimination power would yield only 0.5%. In the case of WFNS, we computed a standardized pAUC of 63.7% with McClish's formula (Equation 1). Of these 63.9%, 50% are due to the small portion (0.5% non-standardized) of the ROC curve below the identity line, and the remaining 13.9% are made of the larger part (2.6% non-standardized) above the curve. In the R version of *pROC*, the standardized pAUC of WFNS can be computed with:

```
roc(response = aSAH$outcome, predictor = aSAH$wfns, partial.auc = c(100, 90), partial.auc.correct = TRUE, percent = TRUE)
```

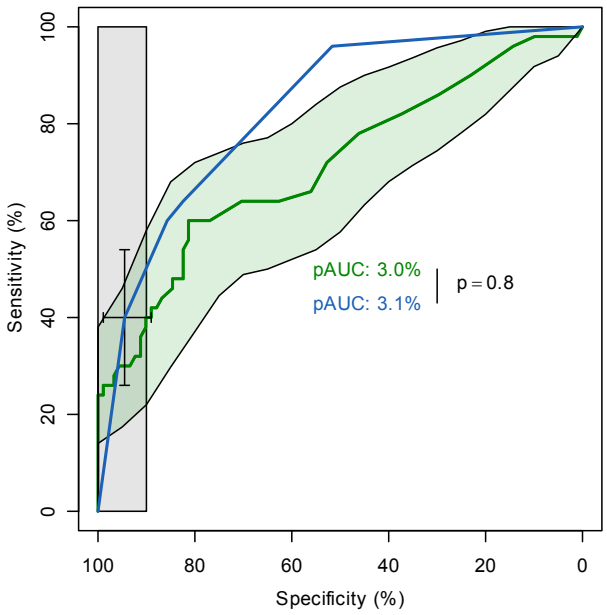


Figure 1 ROC curves of WFNS and S100β. ROC curves of WFNS (blue) and S100β (green). The black bars are the confidence intervals of WFNS for the threshold 4.5 and the light green area is the confidence interval shape of S100β. The vertical light grey shape corresponds to the pAUC region. The pAUC of both empirical curves is printed in the middle of the plot, with the p-value of the difference computed by a bootstrap test on the right.

In the rest of this paper, we report only not standardized pAUCs.

CI

Given the pAUC of WFNS, it makes sense to compute a 95% CI of the pAUC to assess the variability of the measure. In this case, we performed 10000 bootstrap replicates and obtained the 1.6-5.0% interval. In our experience, 10000 replicates give a fair estimate of the second significant digit. A lower number of replicates (for example 2000, the default) gives a good estimate of the first significant digit only. Other confidence intervals can be computed. The threshold with the point farthest to the diagonal line in the specified region was determined with pROC to be 4.5 with the *coords* function. A rectangular confidence interval can be computed and the bounds are 89.0-98.9 in specificity and 26.0-54.0 in sensitivity (Figure 1). If the variability of sensitivity at 90% specificity is considered more relevant than at a specific threshold, the interval of sensitivity is computed as 32.8-68.8. As shown in Figure 1 for S100β, a CI shape can be obtained by simply computing the CI's of the sensitivities over several constantly spaced levels of specificity, and these CI bounds are then joined to generate the shape. The following R code calculates the confidence shape:

```
plot(x = roc(response = aSAH$outcome, predictor = aSAH$s100, percent = TRUE, ci =
```

```
TRUE, of = "se", sp = seq(0, 100, 5)), ci.
type="shape")
```

The confidence intervals of a threshold or of a predefined level of sensitivity or specificity answer different questions. For instance, it would be wrong to compute the CI of the threshold 4.5 and report only the CI bound of sensitivity without reporting the CI bound of specificity as well. Similarly, determining the sensitivity and specificity of the cut-off 4.5 and then computing both CIs separately would also be inaccurate.

Statistical comparison

The second best pAUC is that of S100β with 3.0%. The difference to WFNS is very small and the bootstrap test of *pROC* indicates that it is not significant (p = 0.8, Figure 1). Surprisingly, a Venkatraman's test (over the total ROC curve) indicates a difference in the shape of the ROC curves (p = 0.004), and indeed a test evaluating pAUCs in the high sensitivity region (90-100% sensitivity) would highlight a significant difference (p = 0.005, pAUC = 4.3 and 1.4 for WFNS and S100β respectively). However, since we are not interested in the high sensitivity region of the AUC there is no significant difference between WFNS and S100β.

In *pROC* pairwise comparison of ROC curves is implemented. Multiple testing is not accounted for and in the event of running several tests, the user is reminded that as with any statistical test, multiple tests should be performed with care, and if necessary appropriate corrections should be applied [32].

The bootstrap test can be performed with the following code in R:

```
roc.test(response = aSAH$outcome, predic-
tor1 = aSAH$wfns, predictor2 = aSAH$s100,
partial.auc = c(100, 90), percent = TRUE)
```

Smoothing

Whether or not to smooth a ROC curve is a difficult choice. It can be useful in ROC curves with only few points, in which the trapezoidal rule consistently underestimates the true AUC [17]. This is the case with most clinical scores, such as the WFNS shown in Figure 2 where three smoothing methods available in *pROC* are plotted: (i) normal distribution fitting, (ii) density and (iii) binormal. In our case study:

(i) The normal fitting (red) gives a significantly lower AUC estimate (Δ = -5.1, p = 0.0006, Bootstrap test). This difference is due to the non-normality of WFNS. Distribution fitting can be very powerful when there is a clear knowledge of the underlying distributions, but should be avoided in other contexts.

(ii) The density (green) smoothing also produces a lower (Δ = -1.5, p = $6*10^{-7}$) AUC. It is interesting to note that even with a smaller difference in AUCs, the p-value can be more significant due to a higher covariance.

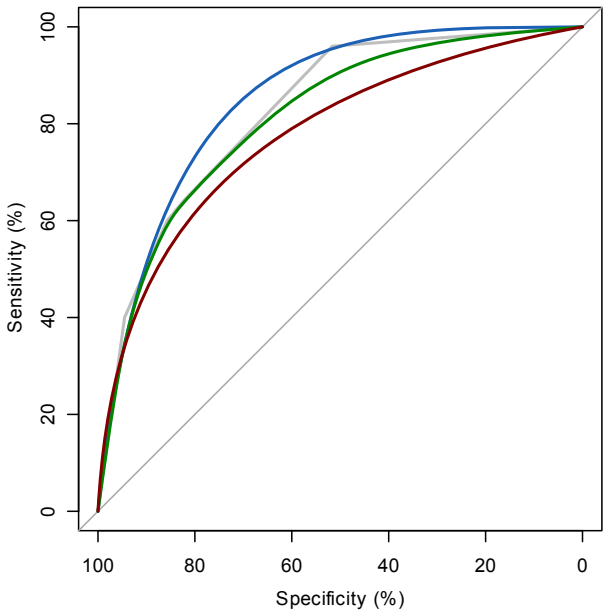


Figure 2 ROC curve of WFNS and smoothing. Empirical ROC curve of WFNS is shown in grey with three smoothing methods: binormal (blue), density (green) and normal distribution fit (red).

(iii) The binormal smoothing (blue) gives a slightly but not significantly higher AUC than the empirical ROC curve (Δ = +2.4, p = 0.3). It is probably the best of the 3 smoothing estimates in this case (as mentioned earlier we were expecting a higher AUC as the empirical AUC of WFNS was underestimated). For comparison, Additional File 5 displays both our implementation of binormal smoothing with the one implemented in pcvsuite [15].

Figure 3 shows how to create a plot with multiple smoothed curves with *pROC* in S+. One loads the pROC library within S+, selects the new *ROC curve* item in the *Statistics* menu, selects the data on which the analysis is to be performed, and then moves to the *Smoothing* tab to set parameters for smoothing.

Conclusion

In this case study we showed how *pROC* could be run for ROC analysis. The main conclusion drawn from this analysis is that none of the measured biomarkers can predict the patient outcome better than the neurological score (WFNS).

Installation and usage

R

pROC can be installed in R by issuing the following command in the prompt:

```
install.packages("pROC")
```

Loading the package:

```
library(pROC)
```

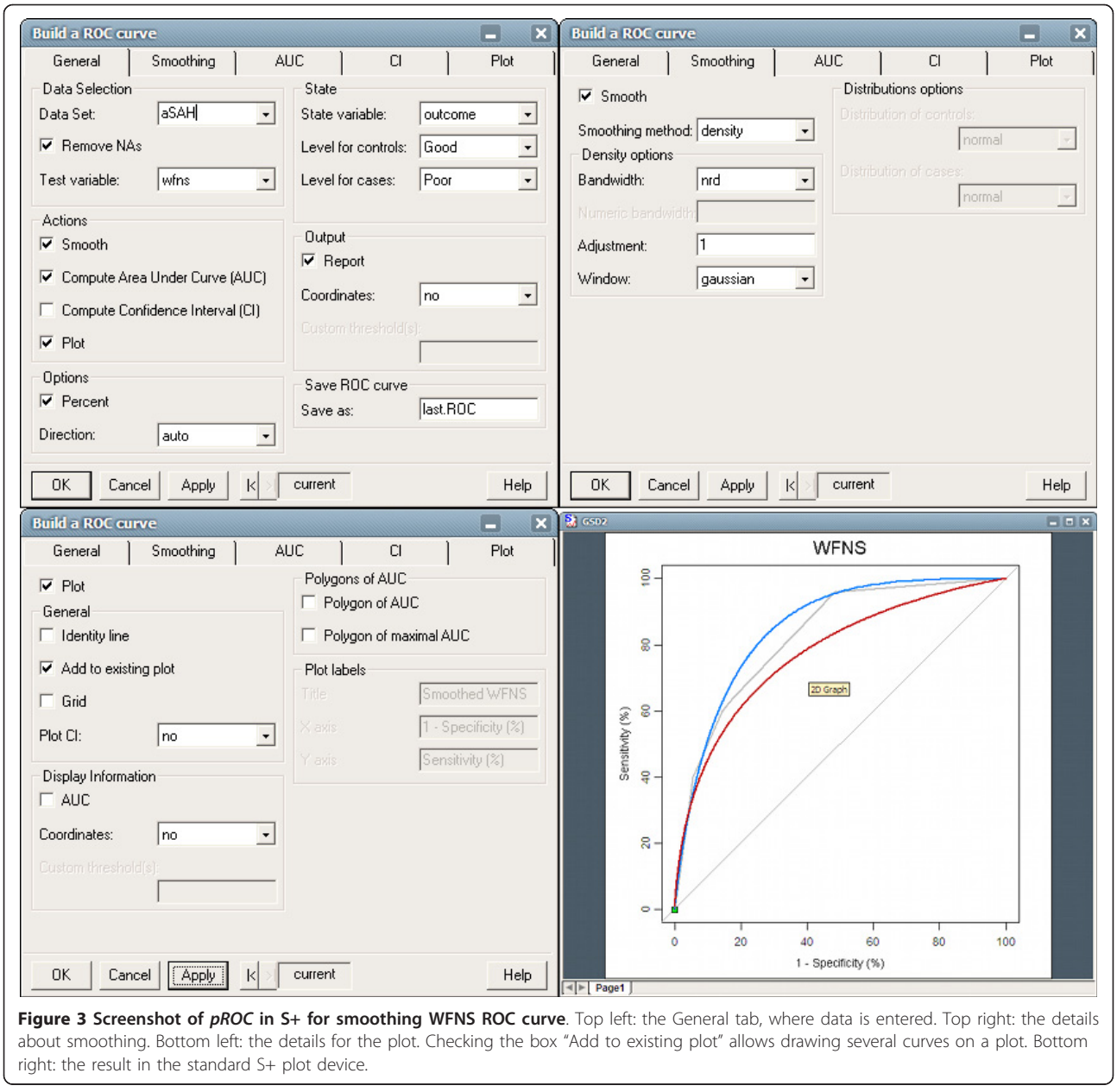


Figure 3 Screenshot of *pROC* in S+ for smoothing WFNS ROC curve. Top left: the General tab, where data is entered. Top right: the details about smoothing. Bottom left: the details for the plot. Checking the box "Add to existing plot" allows drawing several curves on a plot. Bottom right: the result in the standard S+ plot device.

Getting help:

```
?pROC
```

S+

pROC is available from the *File* menu, item *Find Packages....* It can be loaded from the *File* menu, item *Load Library....*

In addition to the command line functions, a GUI is then available in the *Statistics* menu. It features one window for univariate ROC curves (which contains options for smoothing, pAUC, CIs and plotting) and two windows for paired and unpaired tests of two ROC curves. In addition a specific help file for the GUI is available from the same menu.

Functions and methods

A summary of the functions available to the user in the command line version of pROC is shown in Table 2. Table 3 shows the list of the methods provided for plotting and printing.

Conclusions

The *pROC* package is a powerful set of tools analyzing and comparing ROC curves in R and S+. Unlike existing packages such as *ROCR* or *verification*, it is solely dedicated to ROC analysis, but provides in our knowledge the most complete set of statistical tests and plots for ROC curves. As shown in the case study reported here,

Table 2 Functions provided in pROC

are.paired	Determines if two ROC curves are possibly paired
auc	Computes the area under the ROC curve
ci	Computes the confidence interval of a ROC curve
ci.auc	Computes the confidence interval of the AUC
ci.se	Computes the confidence interval of sensitivities at given specificities
ci.sp	Computes the confidence interval of specificities at given sensitivities
ci.thresholds	Computes the confidence interval of thresholds
coords	Returns the coordinates (sensitivities, specificities, thresholds) of a ROC curve
roc	Builds a ROC curve
roc.test	Compares the AUC of two correlated ROC curves
smooth	Smoothes a ROC curve

Table 3 Methods provided by pROC for standard functions

lines	ROC curves (roc) and smoothed ROC curves (smooth.roc)
plot	ROC curves (roc), smoothed ROC curves (smooth.roc) and confidence intervals (ci.se, ci.sp, ci.thresholds)
print	All pROC objects (auc, ci.auc, ci.se, ci.sp, ci.thresholds, roc, smooth.roc)

pROC features the computation of AUC and pAUC, various kinds of confidence intervals, several smoothing methods, and the comparison of two paired or unpaired ROC curves. We believe that *pROC* should provide researchers, especially in the biomarker community, with the necessary tools to better interpret their results in biomarker classification studies.

pROC is available in two versions for R and S+. A thorough documentation with numerous examples is provided in the standard R format. For users unfamiliar with programming, a graphical user interface is provided for S+.

Availability and requirements

- Project name: pROC
- Project home page: http://expasy.org/tools/pROC/
- Operating system(s): Platform independent
- Programming language: R and S+
- Other requirements: R ≥ 2.10.0 or S+ ≥ 8.1.1
- License: GNU GPL
- Any restrictions to use by non-academics: none

Additional material

Additional file 1: Assessment of the ROC comparison tests. We evaluate the uniformity of the tests under the null hypothesis (ROC curves are not different), and the correlation between the different tests.

Additional file 2: Histograms of the frequency of 600 test p-values under the null hypothesis (ROC curves are not different). A: DeLong's paired test, B: DeLong's unpaired test, C: bootstrap paired test (with 10000 replicates), D: bootstrap unpaired test (with 10000 replicates) and E: Venkatraman's test (with 10000 permutations).

Additional file 3: Correlations between DeLong and bootstrap paired tests. X axis: DeLong's test; Y-axis: bootstrap test with number of bootstrap replicates. A: 10, B: 100, C: 1000 and D: 10000.

Additional file 4: Correlation between DeLong and Venkatraman's test. X axis: DeLong's test; Y-axis: Venkatraman's test with 10000 permutations.

Additional file 5: Binormal smoothing. Binormal smoothing with pcvsuite (green, solid) and pROC (black, dashed).

List of abbreviations

aSAH: aneurysmal subarachnoid haemorrhage; AUC: area under the curve; CI: confidence interval; CRAN: comprehensive R archive network; CSAN: comprehensive S-PLUS archive network; pAUC: partial area under the curve; ROC: receiver operating characteristic.

Acknowledgements

The authors would like to thank E. S. Venkatraman and Colin B. Begg for their support in the implementation of their test.
This work was supported by Proteome Science Plc.

Author details

[1]Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland. [2]Swiss Institute of Bioinformatics, Medical University Centre, Geneva, Switzerland.

Authors' contributions

XR carried out the programming and software design and drafted the manuscript. NTu, AH, NTi provided data and biological knowledge, tested and critically reviewed the software and the manuscript. FL helped to draft and to critically improve the manuscript. JCS conceived the biomarker study, participated in its design and coordination, and helped to draft the manuscript. MM participated in the design and coordination of the bioinformatics part of the study, participated in the programming and software design and helped to draft the manuscript. All authors read and approved the final manuscript.

Received: 10 September 2010 Accepted: 17 March 2011
Published: 17 March 2011

References

1. Swets JA: **The Relative Operating Characteristic in Psychology.** *Science* 1973, **182**:990-1000.
2. Pepe MS: *The statistical evaluation of medical tests for classification and prediction* Oxford: Oxford University Press; 2003.
3. Sonego P, Kocsor A, Pongor S: **ROC analysis: applications to the classification of biological sequences and 3D structures.** *Brief Bioinform* 2008, **9**:198-209.
4. Fawcett T: **An introduction to ROC analysis.** *Pattern Recogn Lett* 2006, **27**:861-874.
5. Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER: **Small-sample precision of ROC-related estimates.** *Bioinformatics* 2010, **26**:822-830.
6. Robin X, Turck N, Hainard A, Lisacek F, Sanchez JC, Müller M: **Bioinformatics for protein biomarker panel classification: What is needed to bring biomarker panels into in vitro diagnostics?** *Expert Rev Proteomics* 2009, **6**:675-689.
7. McClish DK: **Analyzing a Portion of the ROC Curve.** *Med Decis Making* 1989, **9**:190-195.
8. Jiang Y, Metz CE, Nishikawa RM: **A receiver operating characteristic partial area index for highly sensitive diagnostic tests.** *Radiology* 1996, **201**:745-750.
9. Streiner DL, Cairney J: **What's under the ROC? An introduction to receiver operating characteristics curves.** *Canadian Journal of Psychiatry Revue Canadienne De Psychiatrie* 2007, **52**:121-128.

10. Stephan C, Wesseling S, Schink T, Jung K: **Comparison of Eight Computer Programs for Receiver-Operating Characteristic Analysis.** *Clin Chem* 2003, **49**:433-439.
11. R Development Core Team: *R: A Language and Environment for Statistical Computing* Vienna, Austria: R Foundation for Statistical Computing; 2010.
12. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R.** *Bioinformatics* 2005, **21**:3940-3941.
13. NCAR: **verification: Forecast verification utilities v. 1.31.**[http://CRAN.R-project.org/package=verification].
14. Carey V, Redestig H: **ROC: utilities for ROC, with uarray focus, v. 1.24.0.** [http://www.bioconductor.org].
15. Pepe M, Longton G, Janes H: **Estimation and Comparison of Receiver Operating Characteristic Curves.** *The Stata journal* 2009, **9**:1.
16. Hanley JA, McNeil BJ: **A method of comparing the areas under receiver operating characteristic curves derived from the same cases.** *Radiology* 1983, **148**:839-843.
17. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.** *Biometrics* 1988, **44**:837-845.
18. Bandos AI, Rockette HE, Gur D: **A permutation test sensitive to differences in areas for comparing ROC curves from a paired design.** *Stat Med* 2005, **24**:2873-2893.
19. Braun TM, Alonzo TA: **A modified sign test for comparing paired ROC curves.** *Biostat* 2008, **9**:364-372.
20. Venkatraman ES, Begg CB: **A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment.** *Biometrika* 1996, **83**:835-848.
21. Bandos AI, Rockette HE, Gur D: **A Permutation Test for Comparing ROC Curves in Multireader Studies: A Multi-reader ROC, Permutation Test.** *Acad Radiol* 2006, **13**:414-420.
22. Moise A, Clement B, Raissis M: **A test for crossing receiver operating characteristic (roc) curves.** *Communications in Statistics - Theory and Methods* 1988, **17**:1985-2003.
23. Venkatraman ES: **A Permutation Test to Compare Receiver Operating Characteristic Curves.** *Biometrics* 2000, **56**:1134-1138.
24. Campbell G: **Advances in statistical methodology for the evaluation of diagnostic and laboratory tests.** *Stat Med* 1994, **13**:499-508.
25. Wickham H: **plyr: Tools for splitting, applying and combining data v. 1.4.** [http://CRAN.R-project.org/package=plyr].
26. Carpenter J, Bithell J: **Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians.** *Stat Med* 2000, **19**:1141-1164.
27. Metz CE, Herman BA, Shen JH: **Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data.** *Stat Med* 1998, **17**:1033-1053.
28. Hanley JA: **The robustness of the "binormal" assumptions used in fitting ROC curves.** *Med Decis Making* 1988, **8**:197-203.
29. Zou KH, Hall WJ, Shapiro DE: **Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests.** *Stat Med* 1997, **16**:2143-2156.
30. Venables WN, Ripley BD: *Modern Applied Statistics with S.* Fourth edition. New York: Springer; 2002.
31. Turck N, Vutskits L, Sanchez-Pena P, Robin X, Hainard A, Gex-Fabry M, Fouda C, Bassem H, Mueller M, Lisacek F, *et al*: **A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage.** *Intensive Care Med* 2010, **36**:107-115.
32. Ewens WJ, Grant GR: **Statistics (i): An Introduction to Statistical Inference.** *Statistical methods in bioinformatics* New York: Springer-Verlag; 2005.

doi:10.1186/1471-2105-12-77
Cite this article as: Robin *et al.*: pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011 **12**:77.

4

PanelomiX: a web-based tool to create biomarker panels based on thresholds

This chapter describes the second tool developed in this thesis, PanelomiX. We present both the tool and the algorithms in detail. It works by combining biomarkers and clinical scores by selecting concentration thresholds that provide optimal classification performance with the iterative combination of biomarkers and thresholds (ICBT) method. Feature selection is carried out with Random Forest when necessary. The robustness and performance of the obtained panels is analyzed with cross-validation and ROC analysis.

We show how this method performs in comparison with separate biomarkers and classical combination methods in the determination of outcome after aneurysmal subarachnoid hemorrhage with 8 parameters on a previously published dataset of 113 patients. The panel classifies the patients better than the best parameter ($p < 0.005$). It also compares favorably with classical methods.

PanelomiX is a tool that allows to combine biomarkers with the ICBT method, and to analyse the robustness and performance of the panels with cross-validation and ROC analysis. ICBT was found to be an efficient and transparent approach to create panels. For the prediction of outcome after aneurysmal subarachnoid haemorrhage, we found a panel comprising 8 parameters and thresholds that could efficiently improve patients' classification in comparison with the individual biomarkers.

This article presents the main results of my thesis. I fully conducted the programming, data analysis, statistics and writing of the manuscript, with input from the other co-authors.

PanelomiX: a web-based tool to create panels of biomarkers based on thresholds

Xavier Robin[a], Natacha Turck[a], Alexandre Hainard[a], Natalia Tiberti[a], Frédérique Lisacek[b], Jean-Charles Sanchez[a] and Markus Müller[b]

[a]Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, University of Geneva, Switzerland
[b]Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland

Abstract

Background: In order to increase their predictive power, medical biomarkers can be combined into panels. However, the lack of ready-to-use tools, to obtain interpretable results with rigorous validation, hampers the more widespread application of panels and their translation into clinical practice.
Methods: The algorithms and platform we present here, called PanelomiX, combines biomarkers and clinical scores by selecting concentration thresholds that provide optimal classification performance with a new method called iterative combination of biomarkers and thresholds (ICBT). Feature filtering is carried out with Random Forest when necessary. The robustness and performance of the obtained panels is analyzed with cross-validation and ROC analysis.
Results: We show how this method performs in comparison with separate biomarkers and classical combination methods for the determination of outcome after aneurysmal subarachnoid haemorrhage with 8 parameters on a previously published dataset of 113 patients. The panel classifies the patients better than the best biomarker ($p < 0.005$). It also compares favourably well with classical methods.
Conclusions: PanelomiX is a tool that allows to combine biomarkers with the ICBT method, and to analyse the robustness and performance of the panels with cross-validation and ROC analysis. ICBT was found to be an efficient and transparent approach to create panels. For the prediction of outcome after aneurysmal subarachnoid haemorrhage, we found a panel comprising 8 biomarkers and thresholds that could efficiently improve patients' classification in comparison with individual biomarkers.

Citation: Robin X., Turck N., Hainard A., Tiberti N., Lisacek F., Sanchez J.-C. and Müller M. PanelomiX: a web-based tool to create panels of biomarkers based on thresholds. Manuscript in preparation.

Abbreviations: ROC: receiver operating characteristic; AUC: area under the ROC curve; pAUC: partial AUC; CV: cross-validation; SE: sensitivity; SP: specificity; aSAH: aneurysmal subarachnoid haemorrhage; SVM: support vector machines; Rpart: recursive partitioning.

Version: 20 August 2012.

Corresponding author. Translational Biomarker Group (9052B), Department of Human Protein Sciences, Centre Medical Universitaire , 1 rue Michel-Servet, CH-1211 Geneva 4, Switzerland . Phone: +41 22 379.54.86. E-mail address: Jean-Charles.Sanchez@unige.ch

Background

The translation of biomarkers or combination of biomarkers into clinical practice is prevented by a number of critical factors[1]. First, methods and results can often be difficult to understand by non-experts; secondly, there is a general lack of robust validation steps, which are critical to ensure reproducible results given the high biological variation and the lack of reproducibility of most experimental methods.

To overcome the former issue, a combination method must present clear and easily interpretable results. This is in opposition with methods such as neural networks or support vector machines (SVM), which may display high classification accuracy when carefully tuned by trained experts, but their inner workings are difficult to understand by end-users who are not experts in statistical learning. While black boxes are acceptable in some specific applications, they may not always be suitable in expert systems for medical decision making[2,3,4]. In contrast, many methods present results in a user-friendly format and can be deemed as white-boxes. Medical practitioners have long been used to clinical scores such as the Hoffer-Osmond test to diagnose schizophrenia[5,6], or the Ranson score[7] for the prognosis and operative management of acute pancreatitis. This kind of scores have become popular because they are clear and easy to interpret, granting access to the intermediate results of individual sub-tests. It was recently applied to assess the probability of pulmonary embolism[8] and acute pancreatitis[9].

Combining biomarkers is an application of statistical learning. Over the years, this field has developed countless methods to tackle this task. Linear or logistic regression methods determine a factor, generally multiplicative, for each biomarker included in the panel. A straightforward interpretation of these factors is to see them as "weights" that increase the importance of the biomarkers having the highest factors. Methods based on decision trees also provide an easy interpretation, following a sequence of single binary biomarker splits on cut-off of the biomarker. As long as the tree contains only a fairly limited number of such decisions (or branches), these are easy to track allowing understanding exactly what is happening and why the decision was reached. Decision trees can be represented graphically[1] for easier understanding. Finally, the threshold-based method is a simplified version of the decision tree. All single biomarker tests are analyzed at the same time (instead of sequentially), and the number of positive tests represents a score, which is used for classification.

The second issue is most of the time associated with the lack of a robust validation step. Validation of a biomarker or a panel requires the availability of an independent test set to compute the true performance of the panel, avoiding the performance overestimation due to the over-fitting of the data during the learning process[1]. If an independent set is not available, computational methods such as cross-validation or bootstrap allow the generation of such sets[10,11].

Useful measures to evaluate the performance are sensitivity (the proportion of positive patients correctly detected by the test) and specificity (the proportion of negative patients correctly rejected by the test), as they give clear information about how well the patients are classified[1]. When no biomarker level cut-off is preferred or pre-defined, receiver operating characteristic (ROC) analysis can be performed to weight the trade-off between sensitivity and specificity[10]. The area under the ROC

curve (AUC) is also a very common performance metric in medical decision making[12], bioinformatics[13] and statistical learning[14]. Finally, an important and often neglected step is to compare the performance of the panel with single biomarkers. A fair comparison would evaluate the panel and single biomarkers with the same tools (sensitivity and specificity or AUC) on the same independent test set or with the same cross-validation procedure[1]. The performance can then be compared either with McNemar's test (for sensitivity or specificity) or through ROC curves.

In threshold-based combinations, thresholds are often chosen in a univariate manner. For example, Ranson[7] selected convenient prognostic sign cut-off values outside the range of the mean plus or minus one standard deviation. Morrow and Braunwald[15] chose the 99th percentile of the control distribution. Sabatine[16] used the cut-offs that were described in the literature. In contrast, Reynolds[17] adopted a multivariate approach and tested many thresholds by 10% increments. This approach takes into account the interaction that may arise when the biomarkers are combined. In PanelomiX we also test the thresholds in a multivariate manner. Unlike the 10% increments adopted by Reynolds[17], the set of cut-offs to be tested is selected from the local maximas on the ROC curve. This guarantees an optimal classification, and is more parsimonious with the non normally distributed data commonly found in clinical studies, where 10% increments of the highest values may not be as significant as those of the lowest ones. To minimize execution times, we developed several approaches to reduce the complexity and hence increase the speed of the search. As it has been shown to be an efficient feature selection method[11], we used Random Forest[18,19] as a filtering method to reduce both the number of biomarkers and of thresholds that accounts for the search space size.

The platform we propose here is called PanelomiX. It can combine biomarkers (molecule levels, clinical scores, etc.) with the threshold-based method, implemented with an exhaustive search algorithm called iterative combination of biomarkers and thresholds (ICBT). Results can be analyzed graphically, and the statistical comparison is performed with full or partial area under the ROC curve. This method has already been applied to predict the outcome of aneurysmal subarachnoid haemorrhage[20] and to assess the progression of human African trypanosomiasis[21].

Methods

ICBT

Combining biomarkers

The method presented here is called ICBT. A threshold is defined for each biomarker by an optimization procedure that will be defined in the next sections. The score of the patient is the count of the biomarkers exceeding their threshold values.

We can write this score as:

$$S_p = \sum_{i=1}^{n} I\left(X_{ip} > T_i\right) , \qquad \text{Equation 1}$$

where S_p is the score for patient p, n is the number of biomarkers, X_{ip} is the concentration of the ith biomarker in patient p, T_i is the threshold for the ith

biomarker and I(x), an indicator function, which takes the value of 1 for x = true and 0 otherwise.

In the case of biomarker concentrations being higher in the control group than in the disease group, it is transformed by taking its opposite before applying the previous formula.

To determine the classification of a patient, a threshold on the score S_p is required and is designed T_s. Patients with a score $S_p \geq T_s$ are positive, negative otherwise.

Choosing the biomarker thresholds

The list of thresholds that will be tested in the ICBT search must be kept short to achieve low computation times. Candidate thresholds are selected as local maximas of the ROC curve, computed with pROC[22]. A local maxima is defined as a point of locally maximal distance to the diagonal line. By definition, we sort the list of biomarkers, resulting in a list of increasing specificity (SP) and decreasing sensitivity (SE). The threshold T_i is a is a local maximum if SP[i] ≥ SP[i-1] and SE[i] ≥ SE[i+1]. Thresholds which do not meet both criteria will not provide a better optimization in a panel because no patient is better classified.

Optimizing the panel

The thresholds are optimized with an exhaustive search. The combinatorial complexity of testing all combinations of biomarkers and threshold values with ICBT can be calculated. Given n biomarkers, and panels with up to m biomarkers, the number C of combinations to test is given by:

$$C = \sum_{i=1}^{m} \binom{n}{i} = \sum_{i=1}^{m} \frac{n!}{i!(n-i)!} \,. \qquad \text{Equation 2}$$

If there are t thresholds per biomarker:

$$I = \sum_{i=1}^{m} \left(\frac{n!}{i!(n-i)!} t^i \right) . \qquad \text{Equation 3}$$

In a typical setup, one would test combinations of 5 or less biomarkers among 10, with 15 thresholds per biomarker. This corresponds to 637 combinations of biomarkers to test. Counting the combination of the thresholds, it sums up to 202.409.025, which is still manageable with current desktop computers.

In most real applications, however, each biomarkers will have a different number of thresholds. If T is a vector containing the number of thresholds of all the biomarkers in combination j, a more precise estimate is given by:

$$I = \sum_{j=1}^{C} \left(\prod T_j \right) . \qquad \text{Equation 4}$$

Pre-filtering

When the computation time becomes too long, an additional step is necessary to reduce the number of biomarker thresholds. Multivariate filtering allows selecting the most interesting biomarkers[11]. In PanelomiX, Random Forest[18,19] is employed as a multivariate filter. The main advantage is that it is possible to analyze the trees created during the process to deduce the biomarkers and thresholds appearing most often and that are thus the most interesting for a combination.

PanelomiX proceeds by stepwise elimination. From the N initial biomarkers, P biomarkers are selected, each associated with Q cut-offs. First a random forest with all the N biomarkers is created. The frequency of appearance of each biomarker in tree branches is extracted and the N-1 biomarkers appearing most often are kept. These two steps are repeated until the target number P of biomarkers is attained. Finally, a last random forest is computed with the P remaining biomarkers to determine the Q most frequently appearing thresholds. As each tree of the random forest is computed from a different set of patients, the cut-offs will differ slightly between the trees of the forest. To be more informative, the bootstrapped thresholds are therefore mapped to the original ones with Euclidean distance. Thresholds are then sorted by frequency and the Q first thresholds of each biomarker are selected for exhaustive searches.

Programming optimizations

At the programming level, several optimizations were implemented to accelerate the ICBT search. First, the compiled language Java was preferred over interpreted languages such as R, Perl or Python, which typically run much slower. Even though Java is an object-oriented language, object-oriented programming has a significant computational cost induced by the creation of multiple objects. Therefore the core of the search program does not create any object and uses only little object-orientation. To make use of all the cores available on the computer that runs the program, multi-threading was implemented with Java's *Runnable* interface.

If the choice of a programming language is extremely important for the sake of execution speed, the algorithmic level has an even stronger effect. We will now discuss three examples.

Firstly, recursive programming functions allow testing panels of an arbitrary size but execute much slower than conventional loops. Therefore PanelomiX is implemented with programmatic loops rather than recursion.

Secondly, instead of computing sensitivity and specificity at each iteration of the algorithm, a pre-defined number of false positives is fixed at the beginning of the search, derived from the target specificity and the number of patients. Only the number of false positives is computed; the panel is further investigated only if it does not exceed this number. If the false positive count is lower, the sensitivity is computed and compared to that of panels previously found. Obviously this can be done only if the panel is set to optimize sensitivity at a minimal level of specificity or conversely specificity at a minimal level of sensitivity. Optimizing the overall accuracy does not allow computing a minimal number of false positives, as a lower count could be acceptable if the number of false negatives is low enough. Therefore panels optimizing the overall accuracy run significantly slower.

Finally, a list of panels is recorded containing all panels with maximal performance. The list is cleared when a panel is found with a higher performance.

The selection of the biomarkers included in the panel is part of the algorithms, therefore it constitutes an embedded feature selection as defined by Dziuda et *al.*[11].

Biomarkers with missing values are ignored. Should the user wish to perform missing values imputation, it must be performed before submitting the data to PanelomiX (see[23] for an in-depth review of the topic).

Cross-validation

Cross-validation is a simple and widely used computational method to assess the performance and the robustness of a classification model[1,10]. PanelomiX features a cross-validation procedure for panel verification[10]. The primary goal of cross-validation is to test its performance in an unbiased manner and to produce graphical diagnostic plots for evaluating its consistency and robustness.

PanelomiX generates several plots to assess the quality of the data. ROC curves analyses are performed on the individual biomarkers and the panel. Several plots are available only after cross-validation.

Centering

To make the predictions comparable between the cross-validation steps which may produce panels of different length and with different T_s, they are centred as follows: first T_s is subtracted from the patient score S:

$$Y_p = S_p - T_s \,. \qquad \text{Equation 5}$$

Then the following transformation is applied:

$$Z_p(Y_p) = \begin{cases} Y_p / T_s, Y_p < 0 \\ Y_p / (n - T_s), Y_p > 0 \end{cases} . \qquad \text{Equation 6}$$

As a result, the centred vector Z of patient scores is comprised in the interval [-1;+1] and T_s=0.

ROC curves

PanelomiX performs and shows the ROC curves of both the individual biomarkers and the panel using the pROC tool[22]. In addition, a table reporting the best thresholds with confidence intervals, and the comparison of the panel with the best biomarker (analyzed as panel composed of 1 biomarker to be comparable with the panels) is generated. Comparisons between two AUCs are performed with DeLong's test[24], and between two pAUCs with the bootstrap test[22] with 10000 stratified replicates. The ROC curves of the cross-validation are built as the mean of centered predictions (equations 5 and 6) over the k CV folds. For the cross-validation of the individual biomarkers, the ICBT algorithm is applied with n = 1 and no other modification.

Implementation

The ICBT search is implemented in the Java programming language, while the other algorithms described above were implemented in R[25]. A perl CGI web interface is also available.

Case study

Patients

The PanelomiX methodology was applied to a previously published dataset of 113 patients with aneurysmal subarachnoid haemorrhage (aSAH). The goal was to identify patients at risk of poor outcome six months after aSAH, who require

specific healthcare management. Detailed results of the study are reported in Turck *et al.*[20]. We will only outline the features relevant to panel analysis.

Panel analysis

Panels were generated as described with five biomarkers (H-FABP, S100-B, Troponin I, NKDA and UFD-1) and three clinical factors (WFNS, Modified Fisher score and age). Cross-validation was performed to assess the performance of the biomarkers, the panels and their stability.

Comparison with standard methods

The results obtained with ICBT were compared with other methods: logistic regression with the *glm* and *step*-wise elimination functions, support vector machines (SVM) from the *kernlab* package[26] (nu-regression with linear kernel), and decision trees from the *rpart* package[27,28]. To be consistent with the ICBT method, both SVM and decision tree feature sets were determined with exhaustive search of all possible combinations. In addition, the predictions were centered as described in equations 5 and 6.

ROC sample size computations

Sample size for the comparison of two ROC curves was implemented according to Obuchowski and McClish [29] (equation 2). To avoid parametric hypotheses about the binormal distribution of the data, variances and covariances of the ROC curves were computed with bootstrap[30].

Results and discussion

Training the panels

The ICBT algorithm was applied on the 113-patients cohort of the aneurysmal subarachnoid hemorrhage study[20]. Combinations of the 8 biomarkers were tested. The optimization criterion was the best accuracy. On this cohort taken as a training set, a panel containing the 8 biomarkers, i.e. the 5 proteins and the 3 clinical parameters was found with the thresholds given in table 1.

Biomarker	H-FABP	S100b	Troponin I	NDKA	UFD-1	WFNS	Age	Fisher Score
Threshold	1.11	0.51	2.33	11.08	271.48	1.5	72.5	2.5
Unit	µg/l	µg/l	µg/l	µg/l	µg/l	N/A	Years	N/A

Table 1: Biomarkers and thresholds in the panel

The performance of the panel was evaluated with two methods: sensitivity and specificity of a threshold, and area under the ROC curve (AUC). On the training set this panel showed 95% sensitivity and 90% specificity, corresponding to an AUC of 95%.

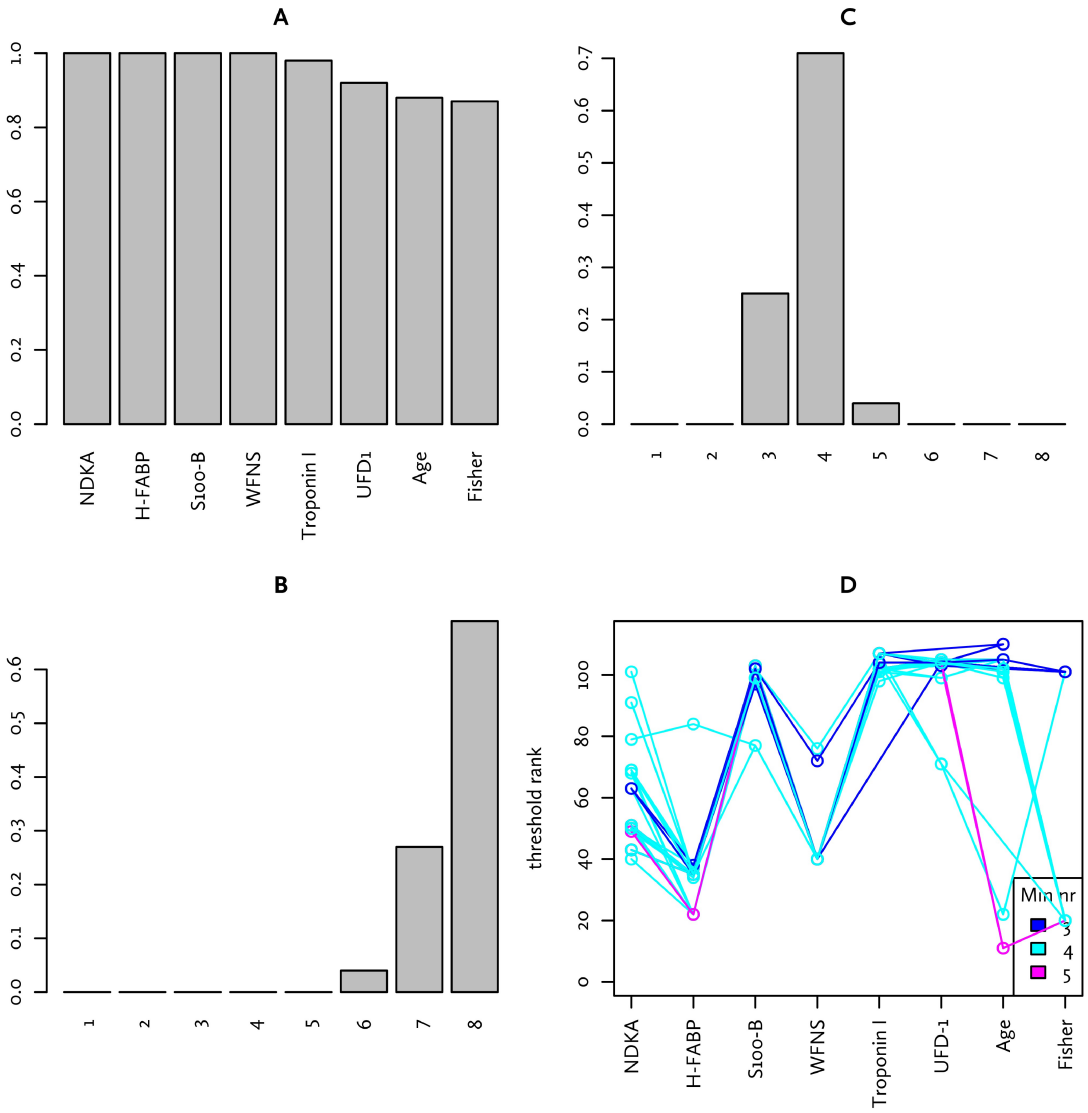


Figure 1: cross-validation plots. A: marker selection frequency plot; B: panel size frequency plot; C: panel Ts frequency plot; D: threshold stability plot.

Cross-validation

Ten-fold cross-validation was repeated 10 times. Four plots that allowed evaluating the stability of the panel with cross-validation are shown in figure 1.

- The *marker selection frequency* plot (figure 1A) shows the frequency of selection of each variable biomarker in the panels trained in the k CV folds. A biomarker with a 100% frequency is selected in all panels. The frequency is weighted: if one step of the cross-validation yields several panels each of them contributes less to the final frequency than panels which are unique in a cross-validation fold. Figure 1A shows that all eight biomarkers selected in the training panel are selected between 88% (Fisher score) and 100% (NDKA, H-FABP, S100b, WFNS) of the cross-validation panels.
- The *panel size frequency* plot displays the number of biomarkers in the panels, weighted as described above. Figure 1B shows that 69% of the cross-validation panels contained 8 biomarkers. In 27% of the panels only 7

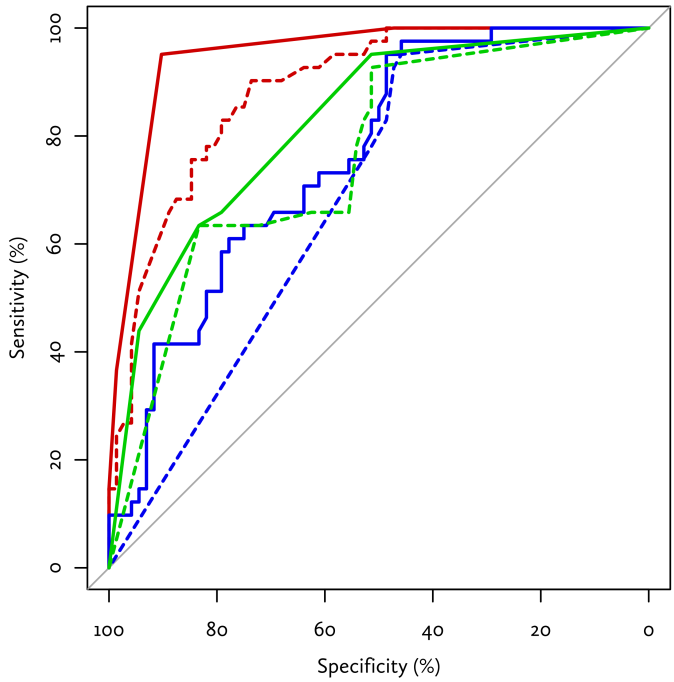


Figure 2: ROC curves. Solid lines represent the performance on the training set, dotted lines the cross-validation. Red: panel of 8 markers; green: WFNS; blue: H-FABP.

biomarkers were selected, and in the 4% remaining 6 biomarkers were selected. No panel containing 5 or less biomarkers was encountered during the cross-validation.

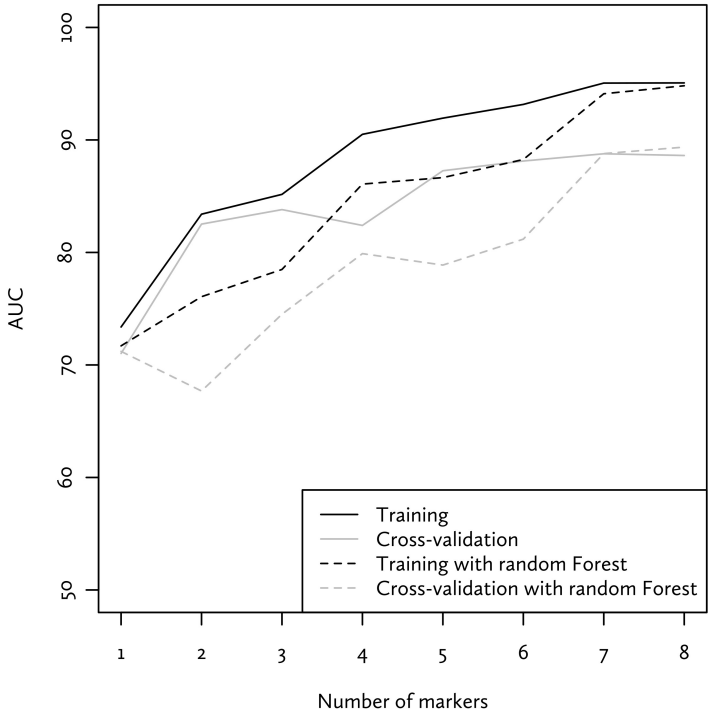


Figure 3: performance of training (black) and cross-validation (grey).

- The *panel* T_s *frequency* plot in figure 1C shows the score T_s, determining how many biomarkers must be positive in a patient for the panel to be positive, weighted as described above. In figure 1C, 25% of the panels had $T_s = 3$, 4% $T_s = 5$ and the rest $T_s = 4$.
- The *threshold stability* plot (figure 1D) represents the biomarkers on the x-axis and the thresholds (as patient rank, not absolute value) on the y-axis of all panels found in the cross-validation. Each panel corresponds to a line joining its constituting set of biomarkers and thresholds. The figure shows that S100b has a very stable threshold, unlike NDKA or UFD-1 that showed a larger variation. In H-FABP, 3 clusters appeared, corresponding to thresholds of 0.61µg/l (rank 22), 1.11µg/l (rank 33) and 4.51µg/l (rank 84). This indicates that the cut-off of NDKA at 11.08µg/l found in the training panel is not as robust as the cut-off at 0.51µg/l found for S100b.

Performance evaluation

A ROC analysis is performed as described in the previous sections (figure 2). The panel found on the training set is plotted together with the cross-validation and the separate biomarkers (see next section). On the cross-validation, panels displayed 65.9% sensitivity and 88.9% specificity, corresponding to an AUC of 88.6%.

Figure 3 shows the performance of the ICBT method on the training set and with cross-validation for panels of different sizes. Panels with 7 biomarkers are optimal in cross-validation, with an AUC (88.8%) slightly higher than panels of 8 (88.6%). However the difference is minimal and it is difficult to determine the significance of this change. This indicates that the level of overfitting induced by ICBT is not too high and that the classification with panels is improved over separate biomarkers.

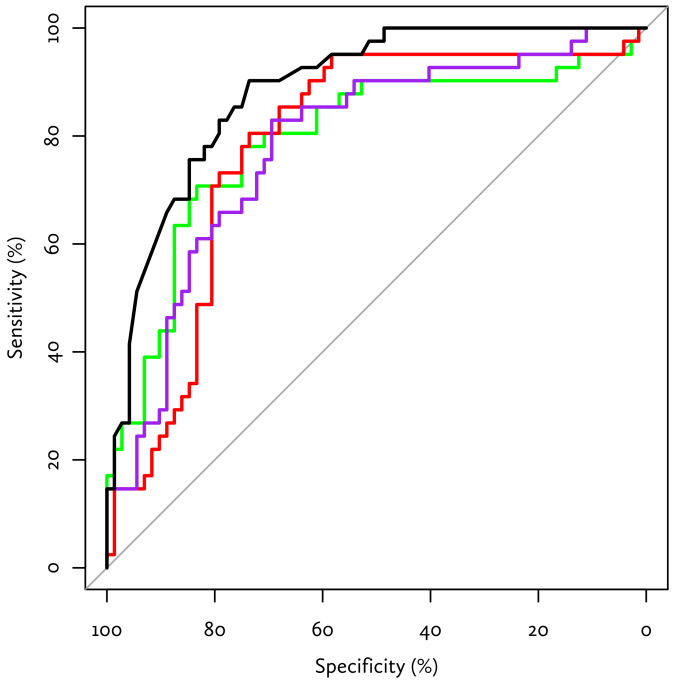


Figure 4: ROC curves showing the comparison with other standard combination methods. Black: PanelomiX; blue: logistic regression; green: SVM; red: Rpart.

Comparison with single biomarkers

Figure 3 shows that, even individual biomarkers are slightly over-fitted and display a lower AUC on the cross-validation (63%) than on the training sample (73%). To perform a fair comparison with ICBT, the cross-validation should not be compared with the biomarkers directly, but with cross-validated biomarkers. To that end, we used the ICBT algorithm where the threshold is chosen on the training set, and applied to the test set.

The two best separate biomarkers, H-FABP and WFNS, are plotted with ICBT in figure 2. The cross-validation (dotted lines) show that panels of 8 biomarkers are superior to the individual biomarkers with an AUC of 89% compared to 76% ($p = 0.003$) for WFNS and 68% ($p = 1.5 \times 10^{-6}$) for H-FABP.

Comparison with established methods

The ICBT method was compared with three established methods among the most widespread in biomarker analysis: logistic regression, SVM and decision trees (recursive partitioning). The results are shown in figure 4. ICBT displayed the best AUC (89%), slightly but not significantly higher than SVM (82%, p=0.20) and logistic regression (81%, p=0.13). Only *rpart* decision trees had a significantly lower AUC of 77% (p=0.03).

Computation time

As stated earlier, the combinations of all 8 biomarkers and all local maxima thresholds can be tested. Table 2 shows the processing time to train a single panel and to perform ten 10-fold cross-validation. The cross-validation of panels of up to 8 biomarkers took slightly less than 6 days to complete on a 4-cores machine.

Size of the panels, *n*	1	2	3	4	5	6	7	8
Training (moria)	0.25	0.34	1.22	8.24	92.92	707.66	2916.43	7082.01
CV (moria)	25.76	32.84	118.57	574.51	5407.57	53415.01	170226.2	380656.37

Table 2: execution time (in seconds) of the panel of increasing size on an Intel Core2 Quad CPU Q9550 at 2.83GHz processor. We show a simple training, and cross-validation (N=10, K=10).

Availability

We built a web interface run PanelomiX from remote computers. It will be made available soon. An package for the R statistical environment is also in preparation.

Conclusions

In this paper we demonstrated that the definition of biomarker panels through exhaustive search is feasible with current computers. Panels created with this methodology are robust and easy to understand even to non-mathematicians. They provide an efficient classification when compared with classical methods. We also proposed several approaches to reduce the complexity and increase the speed of the search for larger setups with only little loss of information.

Finally, we showed how it can be applied to answer to a real clinical question, that is, the outcome prediction ofl patients following aneurysmal subarachnoid haemorrhage. We showed that the PanelomiX algorithm displayed a higher performance compared with classical methods, and that the panel had a superior performance than single biomarkers.

This study suffers from a few limitations. First, only one dataset was analysed in such details. Secondly, no results of Random Forest were reported here. Although we applied these algorithms to datasets with up to a thousand biomarkers (data not shown), more work to improve their robustness is required. Finally, the biomarkers tested here were discovered with univariate approaches. Some of them are relatively highly correlated[20], and multivariate discovery approaches could highlight biomarkers potentially more interesting in combination.

Future prospects include the application of this workflow to datasets with more biomarkers, for instance coming from gene or protein microarrays or single reaction monitoring experiments. It could also be applied to the discovery of new biomarkers with higher classification performance in combination with other biomarkers.

References

1. Robin X., Turck N., Hainard A., et al., (2009). Bioinformatics for protein biomarker panel classification: What is needed to bring biomarker panels into in vitro diagnostics? *Expert Review of Proteomics*, 6 (6), p. 675–689. DOI: 10.1586/EPR.09.83.

2. Duch W., Setiono R. & Zurada J. M., (2004). Computational intelligence methods for rule-based data understanding. *Proceedings of the IEEE*, 92 (5), p. 771– 805. DOI: 10.1109/JPROC.2004.826605.

3. Andrews R., Diederich J. & Tickle A. B., (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8 (6), p. 373–389. DOI: 10.1016/0950-7051(96)81920-4.

4. Baker M., (2005). In biomarkers we trust? *Nat Biotech*, 23 (3), p. 297–304. DOI: 10.1038/nbt0305-297.

5. Hoffer A. & Osmond H., (1961). A card sorting test helpful in making psychiatric diagnosis. *Journal of Neuropsychiatry*, 2, p. 306–330.

6. Kelm H. & Hoffer A., (1965). A revised score for the Hoffer-Osmond diagnostic test. *Diseases of the Nervous System*, 26 (12), p. 790–1.

7. Ranson J. H., Rifkind K. M., Roses D. F., et al., (1974). Prognostic signs and the role of operative management in acute pancreatitis. *Surgery, Gynecology & Obstetrics*, 139 (1), p. 69–81.

8. Wicki J., Perneger T. V., Junod A. F., et al., (2001). Assessing Clinical Probability of Pulmonary Embolism in the Emergency Ward: A Simple Score. *Archives of Internal Medicine*, 161 (1), p. 92–97. DOI: 10.1001/archinte.161.1.92.

9. Imrie C. W., (2003). Prognostic indicators in acute pancreatitis. *Canadian Journal of Gastroenterol*, 17 (5), p. 325–328.

10. Hastie T., Tibshirani R. & Friedman J., (2003). *Elements of Statistical Learning: data mining, inference, and prediction* Springer-Verlag., New York.

11. Dziuda D. M., (2010). *Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data*, John Wiley & Sons.

12. Pepe M. S., (2003). *The statistical evaluation of medical tests for classification and prediction*, Oxford: Oxford University Press.

13. Sonego P., Kocsor A. & Pongor S., (2008). ROC analysis: applications to the classification of biological sequences and 3D structures. *Briefings in Bioinformatics*, 9 (3), p. 198–209. DOI: 10.1093/bib/bbm064.

14. Fawcett T., (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27 (8), p. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

15. Morrow D. A., Rifai N., Antman E. M., et al., (1998). C-Reactive Protein Is a Potent Predictor of Mortality Independently of and in Combination With Troponin T in Acute Coronary Syndromes: A TIMI 11A Substudy. *Journal of the American College of Cardiology*, 31 (7), p. 1460–1465. DOI: 10.1016/S0735-1097(98)00136-3.

16. Sabatine M. S., Morrow D. A., De Lemos J. A., et al., (2002). Multimarker Approach to Risk Stratification in Non-ST Elevation Acute Coronary Syndromes Simultaneous Assessment of Troponin I, C-Reactive Protein, and B-Type Natriuretic Peptide. *Circulation*, 105 (15), p. 1760–1763. DOI: 10.1161/01.CIR.0000015464.18023.0A.

17. Reynolds M. A., Kirchick H. J., Dahlen J. R., et al., (2003). Early Biomarkers of Stroke. *Clinical Chemistry*, 49 (10), p. 1733–1739. DOI: 10.1373/49.10.1733.

18. Breiman L., (2001). Random Forests. *Machine Learning*, 45 (1), p. 5–32. DOI: 10.1023/A:1010933404324.

19. Liaw A. & Wiener M., (2002). Classification and Regression by randomForest. *R News*, 2 (3), p. 18–22.

20. Turck N., Vutskits L., Sanchez-Pena P., et al., (2010). A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Medicine*, 36 (1), p. 107–115. DOI: 10.1007/s00134-009-1641-y.

21. Hainard A., Tiberti N., Robin X., et al., (2009). A Combined CXCL10, CXCL8 and H-FABP Panel for the Staging of Human African Trypanosomiasis Patients. *PLoS Neglected Tropical Diseases*, 3 (6), p. e459. DOI: 10.1371/journal.pntd.0000459.

22. Robin X., Turck N., Hainard A., et al., (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, p. 77. DOI: 10.1186/1471-2105-12-77.

23. Aittokallio T., (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics*, 11 (2), p. 253 –264. DOI: 10.1093/bib/bbp059.

24. Abdi F., Quinn J. F., Jankovic J., et al., (2006). Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders. *Journal of Alzheimer's disease : JAD*, 9 (3), p. 293–348.

25. R Development Core Team, (2008). *R: A Language and Environment for Statistical Computing*, Vienna, Austria.

26. Karatzoglou A., Smola A., Hornik K., et al., (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11 (9), p. 1–20.

27. Therneau T. M. & Atkinson E. J., (1997). *An Introduction to Recursive Partitioning Using the RPART Routines*, Rochester, MN: Mayo Clinic.

28. Therneau T. M., Atkinson B. & Ripley B., (2012). *rpart: Recursive Partitioning*,

29. Obuchowski N. A. & McClish D. K., (1997). Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Statistics in Medicine*, 16 (13), p. 1529–1542. DOI: 10.1002/(SICI)1097-0258(19970715)16:13<1529::AID-SIM565>3.0.CO;2-H.

30. Efron B. & Tibshirani R. J., (1993). *An Introduction to the Bootstrap* Chapman & Hall., New York, London.

5

A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage

This chapter describes a first application of PanelomiX. The purpose of this study was to improve the prediction of long-term irreversible brain damage during the acute phase of patients with aneurysmal subarachnoid hemorrhage (aSAH). A pronostic panel was developed to facilitate early outcome prediction following aSAH, with a combination of clinical scores together with brain injury-related biomarkers.

To this aim, two cohorts of patients (a selection set of 28 patients and a verification set of 113 patients) were prospectively enrolled, and venous blood samples collected within 12 hours after admission and within 48 h following aSAH onset were analyzed. Five proteins, H-FABP, NDKA, UFD1, S100-B and troponin I, were measured with classical immunoassays. Three clinical measurements, the WFNS, a modified Fisher score and age were assessed as biomarkers. The outcome after 6 months was evaluated with the Glasgow Outcome Score (GOS) and used as gold standard. A favorable outcome was defined when a $1 \leq GOS \leq 3$, and an unfavorable outcome was defined when $4 \leq GOS \leq 5$. The classification power of both the biomarkers and the panel was assessed through ROC analysis and partial AUC (pAUC) with pROC.

A panel comprising six biomarkers was found, comprising the WFNS score and the 5 proteins, H-FABP, S100-B, troponin I, NDKA and UFD-1. The panel was positive when at least three of these parameters were simultaneously above the determined cutoff values. On the verification set of 113 patients, the prediction of unfavorable outcome displayed 70% sensitivity and 100% specificity.

In conclusion, this panel, including four brain injury-related proteins, one cardiac marker and a clinical score, could be a valuable tool to identify aSAH patients at risk of poor outcome.

In this article, I contributed to the data analysis and statistics. Specifically, I defined the multiparameter panel and conducted the ROC analysis. I was also involved in the remainder of the statistical analysis.

Intensive Care Med (2010) 36:107–115
DOI 10.1007/s00134-009-1641-y

ORIGINAL

Natacha Turck
Laszlo Vutskits
Paola Sanchez-Pena
Xavier Robin
Alexandre Hainard
Marianne Gex-Fabry
Catherine Fouda
Hadiji Bassem
Markus Mueller
Frédérique Lisacek
Louis Puybasset
Jean-Charles Sanchez

A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage

Received: 28 January 2009
Accepted: 30 July 2009
Published online: 17 September 2009

Electronic supplementary material
The online version of this article (doi:10.1007/s00134-009-1641-y) contains supplementary material, which is available to authorized users.

N. Turck (✉) · X. Robin · A. Hainard · C. Fouda · J.-C. Sanchez
Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, DBSB/CMU, Rue Michel Servet, 1, 1211 Geneva 4, Switzerland
e-mail: natacha.turck@unige.ch
Tel.: +41-22-3795906
Fax: +41-22-3795984

L. Vutskits
Department of Anesthesiology, Pharmacology and Intensive Care, University Hospital of Geneva, 1211 Geneva 14, Switzerland

P. Sanchez-Pena · H. Bassem · L. Puybasset
Department of Anesthesiology and Critical Care, Pitié-Salpêtrière Teaching Hospital, Assistance Publique-Hôpitaux de Paris and Université Pierre et Marie Curie-Paris 6, Paris 75013, France

X. Robin · M. Mueller · F. Lisacek
Swiss Institute of Bioinformatics, Medical University Centre, 1211 Geneva 4, Switzerland

M. Gex-Fabry
Department of Psychiatry, University Hospital of Geneva, 1225 Chêne-Bourg, Switzerland

Abstract *Purpose:* Accurate early anticipation of long-term irreversible brain damage during the acute phase of patients with aneurysmal subarachnoid hemorrhage (aSAH) remains difficult. Using a combination of clinical scores together with brain injury-related biomarkers (H-FABP, NDKA, UFD1 and S100β), this study aimed at developing a multiparameter prognostic panel to facilitate early outcome prediction following aSAH. *Methods:* Blood samples of 141 aSAH patients from two separated cohorts (sets of 28 and 113 patients) were prospectively enrolled and analyzed with 14 months of delay. Patients were admitted within 48 h following aSAH onset. A venous blood sample was withdrawn within 12 h after admission. H-FABP, NDKA, UFD1, S100β and troponin I levels were determined using classical immunoassays. The World Federation of Neurological Surgeons (WFNS) at admission and the Glasgow Outcome Score (GOS) at 6 months were evaluated. *Results:* In the two cohorts, blood concentration of H-FABP, S100β and troponin I at admission significantly predicted unfavorable outcome (GOS 1–2–3). A multivariate analysis identified a six-parameter panel, including WFNS, H-FABP, S100β, troponin I, NDKA and UFD-1; when at least three of these parameters were simultaneously above cutoff values, prediction of unfavorable outcome reached around 70% sensitivity in both cohorts for 100% specificity. *Conclusion:* The use of this panel, including four brain injury-related proteins, one cardiac marker and a clinical score, could be a valuable tool to identify aSAH patients at risk of poor outcome.

Keywords Aneurysmal subarachnoid hemorrhage · H-FABP · NDKA · S100β · Prognosis

Abbreviations

H-FABP	Heart-fatty acid binding protein
NDKA	Nucleoside diphosphate kinase A
UFD-1	Ubiquitin fusion degradation protein-1
aSAH	Aneurysmal subarachnoid hemorrhage
SE	Sensitivity
SP	Specificity

Introduction

Besides the high early mortality associated with aSAH, long-term neurological morbidity is also a significant problem in a substantial proportion of these patients [1, 2]. Identification of prognostic factors, aimed to predict patient outcome, would help in the management and decision making within this population. Clinical scores, such as WFNS classification, demonstrated an association between prognosis following aSAH and the patient's clinical neurological status at hospital admission [3]. Biochemical markers may provide additional information about specific pathological disruptions and recovery processes that occur in the central nervous system following aSAH. In conjunction with clinical status, these biomarkers may also inform prognosis and guide therapeutic decisions to optimize treatments [4, 5].

Over the past few years, a large number of biomarkers, present in the blood and CSF, have raised interest in the detection of aSAH patients with poor clinical outcome. Nevertheless, the majority of these markers displayed either low sensitivity or specificity to anticipate the detection of patients with poor outcome [6, 7].

We recently explored post-mortem CSF as a model of massive brain insult [8, 9]. In these studies, heart-fatty acid binding protein (H-FABP), nucleotide diphosphate kinase A (NDKA) and ubiquitin fusion degradation protein 1 (UFD-1) were over-expressed in post-mortem compared to ante-mortem CSF and were validated as potential brain damaged biomarkers [10–12]. In the present study, we hypothesized that such a reliable plasmatic marker may provide quantitative information reflecting the prediction of aSAH patient outcome. The objective of this study was to determine, immediately at the hospital admission, S100β, H-FABP, troponin I, NDKA and UFD-1 protein blood concentrations of patients with spontaneous aSAH obtained in two separated cohorts from the same institution. In addition to specific clinical parameters, their potential predictive power to detect poor 6-month outcome following aSAH was evaluated [13–15].

Patients and methods

Population

The inclusion period was from July 2004 to December 2006 in the Pitié-Salpêtrière Hospital (Paris, France). Inclusion criteria were clinical history of aSAH within the last 2 days before admission with evidence of bleeding in CT and presence of an aneurysm at cerebral angiography, age above 18 years old and treatment by surgery or coiling within 48 h after admission. Each eligible patient was admitted in the intensive care unit (ICU) within the 2 days after aSAH symptom onset (mean 7 ± 18 h, min 3 h and max 48 h), and a unique venous blood sample was withdrawn within 12 h after ICU admission (mean 24 ± 13.9 h). Fifty-nine patients were excluded due to either a delay of more than 48 h after the onset of symptoms ($n = 55$) or missing clinical information ($n = 3$). A total of 199 consecutive patients were evaluated, and 141 were finally enrolled in this study.

Samples were sent from Paris to Geneva in two distinct sets of samples with a 14-month period delay. As samples were analyzed immediately in Geneva, results between the two sets displayed a 14-month period delay explaining why the two sets were considered separately. The selection set had 28 patients (8 men and 20 women; age range 26–84 years) and the verification set 113 patients (42 men and 71 women; age range 18–81 years). Fifty patients (35.4% of the study sample) had an unfavorable outcome at 6 months (GOS score 1–3), and 91 (65%) patients had a favorable outcome (GOS score 4–5). The two sets are described in Table 1.

The local ethical committee (Comité de Protection des Personnes, Pitié-Salpêtrière, Paris, France) approved the study. In accordance with the Helsinki Declaration, written informed consent was obtained from the patient or patient's relatives.

Clinical monitoring and treatment

At admission, clinical severity was assessed using the WFNS score [16]. The initial CT was reviewed by an independent radiologist blinded to clinical history and classified according to the original Fisher score [17] modified as follows: grade 1, no subarachnoid blood; grade 2, broad diffusion of subarachnoid blood; grade 3, with clots or thick layers of blood; grade 4, intraventricular hemorrhage or intracerebral hematoma, no clot; grade 5, intraventricular hemorrhage or intracerebral hematoma with clot [18–20] and qualified presence or absence of acute hydrocephalus. The neurological outcome was assessed by phone interviews using the Glasgow Outcome Scale (GOS) [21] at 6 months. The type of treatment (surgery or coiling) was decided according to both location and size of the aneurysm by the neurosurgeon and the neuro-radiologist. Seizures were routinely prevented by gabapentin (600 mg t.i.d., per os). A central venous line and an arterial catheter were inserted in most of the patients before and/or after surgery or coiling. An external ventricular drain (Sophysa, Orsay, France) was inserted in patients with CT evidence of hydrocephalus, high WFNS grade or a trans-cranial Doppler (TCD) pulsatility index greater than 1.2, suggesting intracranial pressure (ICP) elevation. The line was

Table 1 Main characteristics of the population

	28-Patient set			113-Patient set		
	GOS 1–2–3 ($N = 9$)	GOS 4–5 ($N = 19$)	P^a	GOS 1–2–3 ($N = 41$)	GOS 4–5 ($N = 72$)	P^a
Gender			1			0.07
♂ *n%*	3 (33.3)	5 (26.3)		20 (48.8)	22(30.6)	
♀ *n%*	6 (66.4)	14 (73.7)		21(51.2)	50(69.4)	
Age (years)			0.86			0.043
Median (range)	56 (49–75)	57 (26–84)		55.0 (31–81)	49.5 (18–76)	
Mean (±SD)	56.9 (±7.4)	53.5 (±14.1)		54.9 (±13.3)	48.9 (±13.8)	
Time of blood drawing (h)			0.74			0.73
Median (range)	24 (6–24)	22.5 (11–48)		24 (10–48)	24 (5–48)	
Mean (±SD)	20.4 (±6.3)	22.9 (±11.8)		21.8 (±10.5)	20.9 (±9.9)	
WFNS score			0.026			<0.0001
1–2 *n%*	4 (44.4)	18 (94.8)		14 (34.1)	57 (79.2)	
3–4–5 *n%*	5 (55.6)	1 (5.2)		27 (67.5)	15 (20.8)	
Modified Fisher score			0.14			<0.0001
1–2 *n%*	0 (0.0)	5 (26.3)		0 (0.0)	19 (26.4)	
3–4–5 *n%*	9 (100.0)	14 (73.8)		41 (100.0)	53 (73.6)	
Vasospasm			0.08			0.48
No *n%*	6 (66.7)	18 (94.8)		27 (65.9)	54 (75.0)	
Yes *n%*	3 (33.3)	1 (5.2)		14 (34.1)	19 (25.0)	
Location			0.23			0.32
MCA *n%*	3 (33.3)	1 (5.2)		12 (29.3)	11 (15.3)	
CA *n%*	3 (33.3)	10 (52.7)		19 (46.3)	36 (50.0)	
ICA/PCA *n%*	3 (33.3)	7 (36.9)		10 (24.4)	23 (31.9)	
VBS *n%*	0	1 (5.2)		0 (0.0)	2 (2.8)	
Treatment			0.12			0.29
No *n%*	1 (11.1)	0		2 (4.9)	2 (2.8)	
Coiling *n%*	6 (66.7)	18 (94.8)		29 (70.7)	60 (83.3)	
Surgery *n%*	2 (22.2)	1 (5.2)		10 (24.4)	10 (13.9)	
Seizures			1			0.89
No *n%*	6 (66.7)	13 (68.4)		33 (80.5)	58 (80.0)	
Yes *n%*	3 (33.3)	6 (31.6)		8 (19.5)	14 (19.4)	

Age non-parametric Mann–Whitney *U* test

MCA middle cerebral artery, *CA* cerebral anterior artery, *ICA* internal carotid artery, *PCA* posterior communicating artery, *VBS* vertebro basilar system

[a] Fisher exact test

connected to an external pressure strain gauge to monitor ICP. Early ICP elevation was defined as ICP above 20 mmHg under sedation but without drainage. Monitoring and treatment of vasospasm are described in Online Data Supplement 1.

H-FABP, S100β, NDKA, UFD1 and troponin I measurements

Cardiac troponin I serum concentration was systematically measured using the Stratus Analyzer (Dade, Massy, France). S100β concentration was measured with an immunoluminometric sandwich assay on a LIA-mat 300 analyzer (Byk-Sangtec France Laboratories, Le Mée sur Seine, France) using the manufacturer's reagents [22]. H-FABP concentration was determined with a commercially available enzyme-linked immunosorbent assay (ELISA) (Hycult Biotechnology, Uden, The Netherlands) according to the manufacturer's instructions. The concentrations of NDKA and UFD1 were determined by home-made ELISA as previously described by Allard et al. [11, 12]. For more details, see the Online Data Supplement 2.

Data analysis and statistics

SPSS software (version 15, SPSS Inc., Chicago, IL), R (URL http://www.R-project.org) and PERL (ActivePerl version 5.8.8.820, ActiveState Software Inc.) were used for data analysis.

Because protein concentrations did not show normal distributions (Kolmogorov–Smirnov test), between-group differences were tested with the non-parametric Mann–Whitney *U* test. The Fisher exact test was used for categorical variables. Statistical significance was set at 0.05 (two-tailed tests).

The dichotomized 6-month GOS score was considered as the main outcome variable, with ranges 1–2–3 and 4–5 reflecting unfavorable and favorable outcome, respectively.

The different markers (H-FABP, S100β, troponin I, NDKA, UFD-1) as well as clinical data were considered as possible predictors.

For each individual predictor, a receiver-operating characteristic (ROC) curve was determined in each cohort, and a cutoff value was selected as the threshold predicting poor outcome with specificity >90%. Partial ROC AUCs (pAUC) [23, 24] and 95% confidence intervals (CI) were calculated using an adaptation of previously described algorithms [25]. pAUCs were restricted between 90 and 100% specificity considering that an efficient predictor in clinical practice should be able to identify clearly at least nine out of ten patients as having a favorable prognosis when the test was negative. *P* values for the difference between two pAUCs were computed based on [26] where standard deviation was determined by bootstrap as described above.

Univariate and multivariate logistic regressions with stepwise backward selection were performed using SPSS software and are described in Online Data Supplements 3 and 4, respectively.

Panel development

Panel selection was performed essentially as described by Reynolds et al. [27, 28]. Briefly, the optimized cutoff values were obtained by iterative permutation-response calculations using all available parameters. Each cutoff value was changed iteratively by quantiles of 2% increment, and sensitivity was determined after each iteration until a maximum of sensitivity was achieved for 100% specificity. Binary clinical parameters (hydrocephaly, vasospasm, sex and statin treatment) were recorded as 0/1 (absent/present), and a unique cutoff of 0.5 was used.

Results

Patients with favorable and unfavorable outcomes did not significantly differ with respect to gender. Age of patients with poor outcome at 6 months was slightly higher in the 113-patient set, suggesting that age might be considered as a prognostic factor. WNFS score was significantly higher in patients with a poor outcome than favorable outcome (Fisher's exact test, $P = 0.026$ and <0.0001 in the 28- and 113-patient sets, respectively). A modified Fisher score, estimating severity of aSAH, did not significantly differ according to outcome in the 28-patient set, whereas in the 113-patient set, severe aSAH (high modified Fisher score) was significantly associated with unfavorable outcome (Fisher's exact test, $P < 0.0001$). No associations were found between long-term neurological outcome and time course of blood samples drawings, post-hemorrhagic seizures, location of the aneurysm, occurrence of vasospasm and treatment modality (coiling vs. surgery). Demographic characteristics are shown in Table 1.

As shown in Table 2, baseline H-FABP, S100β and troponin I levels were significantly elevated in the blood of patients with an unfavorable outcome compared to patients with a favorable outcome. Initial NDKA and UFD-1 levels were unable to discriminate between favorable and unfavorable outcome in the 28-patient set, but in the 113-patient set, the NDKA level was marginally higher in patients with a poor 6-month outcome

Table 2 H-FABP, S100β, troponin I, NDKA and UFD-1 concentrations (µg/l) at admission according to the patient outcome at 6 months in the 28- and 113-patient sets

	28-Patient set			113-Patient set		
	GOS 1–2–3 ($N = 9$)	GOS 4–5 ($N = 19$)	*P*	GOS 1–2–3 ($N = 41$)	GOS 4–5 ($N = 72$)	*P*
H-FABP (µg/l)						
Median (range)	4.65 (1.73–62.2)	1.79 (0.86–9.03)	0.01	3.59 (0.63–67.36)	1.35 (0–44.43)	<0.0001
Mean (±SD)	12.06 (±19.16)	2.82 (±2.13)		10.33 (±16.30)	3.50 (±7.33)	
S100β (µg/l)						
Median (range)	0.33 (0.17–0.46)	0.15 (0.06–0.32)	0.04	0.30 (0.03–2.07)	0.11 (0.04–0.5)	<0.0001
Mean (±SD)	0.32 (±0.14)	0.163 (±0.84)		0.39 (±0.37)	0.16 (±0.13)	
Troponin I (µg/l)						
Median (range)	0.50 (0.04–6.4)	0.05 (0.04–2.62)	0.04	0.36 (0.03–155)	0.05 (0.03–4.4)	<0.0001
Mean (±SD)	1.92 (±2.54)	1.15 (±3.18)		5.51 (±24.2)	0.32 (±0.77)	
NDKA (µg/l)						
Median (range)	13.74 (0–46.39)	13.98 (2.31–32.81)	0.92	13.56 (3.9–419.2)	10.95 (3.0–80.3)	0.05
Mean (±SD)	15.6 (±13.76)	15.9 (±10.45)		28.08 (±64.2)	14.86 (±12.8)	
UFD-1 (µg/l)						
Median (range)	71.0 (1.83–24.55)	12.6 (0.39–33.8)	0.33	83.73 (3.61–1792)	84.48 (10.4–553.2)	0.99
Mean (±SD)	11.23 (±8.14)	15.06 (±10.21)		169.3 (±291.6)	108.3 (±87.9)	

P = Non-parametric Mann–Whitney U test. $P < 0.05$ is considered significant

($P = 0.073$, Mann–Whitney U test). No significant difference was observed in the molecule concentrations as a function of time of blood drawing (data not shown).

The prediction performances of individual molecules, neurological scales and age for predicting a poor outcome were evaluated with ROC curves and pAUC (Fig. 1). Thresholds of individual predictors were chosen to provide specificity above 90% except for WFNS and modified Fisher where the cutoff value was fixed to separate patients according to their clinical pattern. With a threshold strictly above 2, WFNS allowed to discriminate patients with poor and favorable outcome with 55.6% sensitivity (SE) and 97.4% specificity (SP) in the 28-patient set and 67.5% SE for 79.2% SP in the 113-patient set. The modified Fisher scale (threshold >2) provided perfect 100% sensitivity in the two sample sets but low specificity (26.4 vs. 29.0%).

In the 28-patient set, H-FABP, S100β and troponin I displayed 44.4, 33.0 and 22.2% SE for 94.7, 94.0 and 94.7% SP, respectively. Similar performances were obtained in the 113-patient set. NDKA, UFD1 and age led to relatively poor prediction of outcome at 6 months in the two sets (Tables 3, 4). Univariate and multivariate logistic regressions with stepwise backward selection were used to

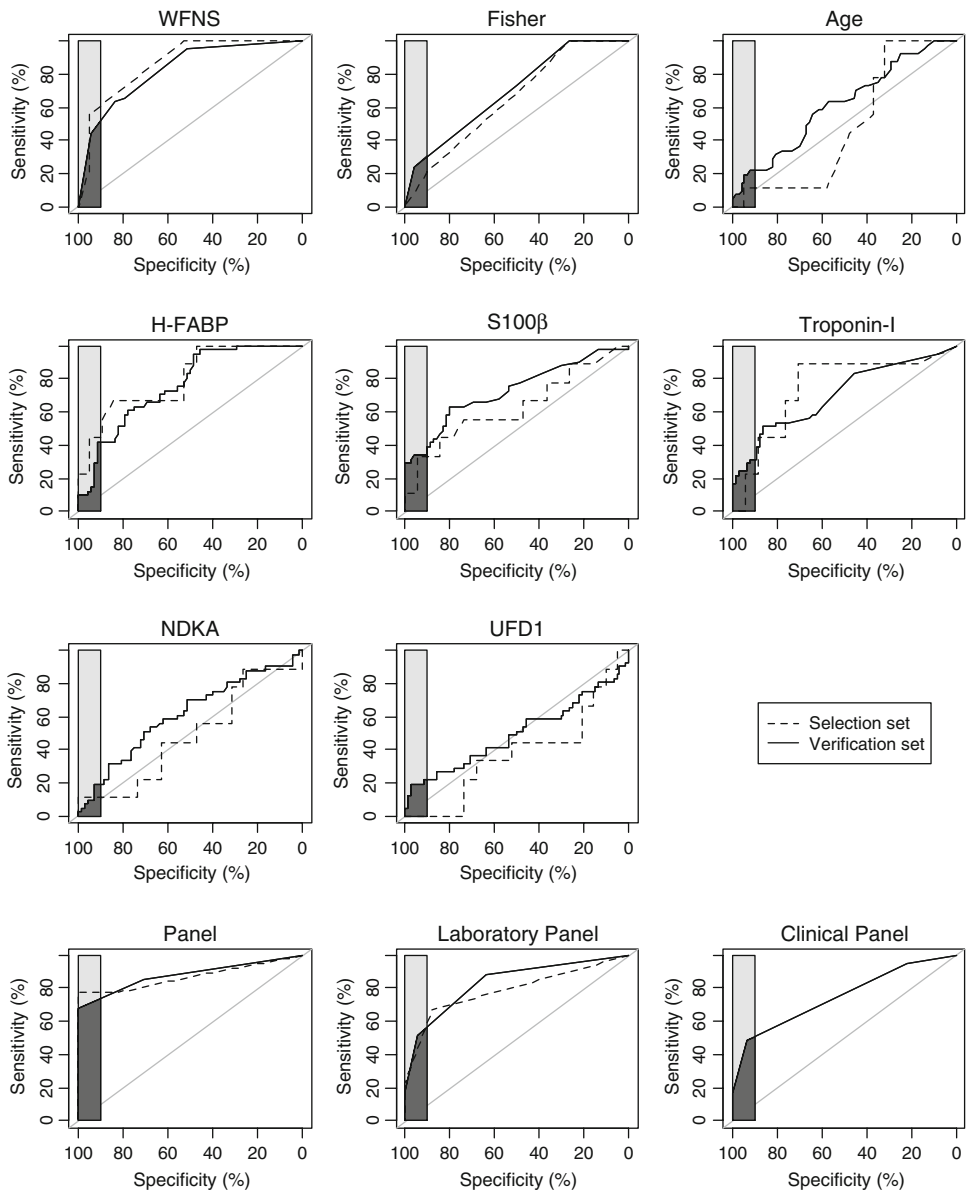


Fig. 1 pAUC of the different parameters on the selection ($N = 28$) and verification ($N = 113$) sets. *Grey boxes* correspond to the maximal area (10%) between 90 and 100% SP if a perfect ROC curve was obtained. *Dark boxes* correspond to the partial area under the curve

Table 3 Partial area under the curve (pAUC), sensitivities (SE) and specificities (SP) for individual parameters and the panel on the 28-patient set

28-Patient set				
	Partial AUC (95% CI)	Threshold	SE (%) (95% CI)	SP (%) (95% CI)
WFNS	3.0% (0.0–8.2)	>2[a]	55.6 (20.0–88.9)	94.1 (81.3–100)
Modified Fisher	1.7% (0.0–3.8)	>2[a]	100 (100–100)	29.4 (8.3–52.9)
Age	1.1% (0.0–3.6)	72.5 years old	11.1 (0.0–33.3)	100 (100–100)
H-FABP	5.1% (1.7–8.8)	6.3 μg/l	44.4 (12.5–80.0)	100 (100–100)
S100β	2.0% (0.0–6.7)	0.37 μg/l	33.3 (0.0–66.7)	94.1 (80.0–100)
Troponin I	0.9% (0.0–6.7)	5.3 μg/l	22.2 (0.0–50.0)	94.1 (80.0–100)
NDKA	1.1% (0.0–3.8)	31.9 μg/l	11.1 (0–37.5)	100 (100–100)
UFD-1	0.0% (0.0–1.4)	24.87 μg/l	0 (0–0)	76.5 (53.8–94.4)

[a] A threshlod strictly above two for the neurological scores means that patients have been dichotomized into two groups: patients with WFNS 1–2 and patients with WFNS 3–5 or patients with modified Fisher 1–2 and patients with modified Fisher 3–5

Table 4 Partial area under the curve (pAUC), sensitivities (SE) and specificities (SP) for individual parameters and the panel on the two sets

113-Patient set				
	Partial AUC (95% CI)	Threshold	SE (%) (95% CI)	SP (%) (95% CI)
WFNS	3.3% (1.7–5.4)	>2[a]	65.9 (51.1–79.2)	79.1 (69.6–88.1)
Modified Fisher	2.1% (0.9–3.6)	>2[a]	100 (100–100)	26.4 (16.7–36.8)
Age	1.5% (0.6–2.6)	67.5 years old	20.4 (9.5–32.6)	92.0 (86.0–97.4)
H-FABP	1.9% (0.5–4.0)	5.9 μg/l	41.4 (26.5–56.8)	91.7 (84.7–97.3)
S100β	3.3% (1.9–4.9)	0.48 μg/l	31.7 (17.8–46.5)	97.2 (92.9–100)
Troponin I	2.5% (1.2–4.3)	1.56 μg/l	29.3 (15.6–43.9)	93.1 (86.8–98.6)
NDKA	1.0% (0.2–2.5)	30.4 μg/l	19.5 (8.3–32.4)	93.1 (86.5–98.6)
UFD-1	1.7% (0.6–3.0)	271.5 μg/l	19.5 (7.9–32.5)	97.2 (93.0– 100)

[a] A threshlod strictly above two for the neurological scores means that patients have been dichotomized in two groups: patients with WFNS 1–2 and patients with WFNS 3–5 or patients with modified Fisher 1–2 and patients with modified Fisher 3–5

validate predictors of poor outcome. Results are presented in Online Data Supplements 3 and 4, respectively.

Provided the low sensitivity obtained with individual predictors, we tested combinations of all parameters on the 28-patient set to select a panel that could improve outcome prediction. The iterative permutation-response approach led to a six-parameter panel including WFNS, H-FABP, S100β, troponin I, NDKA and UFD-1. The panel result was defined as positive if at least three out of the six selected parameters were simultaneously above threshold with 77% (95% CI: 50.0–100.0%) SE for 100% (95% CI: 100.0–100.0) SP. This panel tested on the 113-patient set presented extremely similar performances with 68.3% (95% CI: 53.5–82.2) SE for 100% (95% CI: 100.0–100.0) SP. The panel was confirmed by a ten-fold cross-validation (data not shown). The six-parameter panel allowed to increase sensitivity by about 25% when compared to the best single predictor (H-FABP, SE: 45%). In addition, pAUC of the panel was significantly higher than pAUC of WFNS ($P < 0.0002$). The relative performance of each marker was also evaluated by removing them one by one from the panel and recalculating sensitivity and specificity. The results obtained are shown in Online Data Supplement 5.

Importantly including both clinical and laboratory variables into the same panel was found to be superior to approaches combining either purely clinical or merely laboratory variables. Indeed, a clinical panel including WFNS, modified Fisher scale and age displayed only 22% SE for 100% SP to predict poor outcome in the 113-patient set. The presence of hydrocephaly and occurrence of vasospasm in the clinical panel did not improve the discriminating performance of the clinical panel. In line with this, a panel containing only the five laboratory variables reached only 50% SE for 100% SP (Fig. 1). pAUC of our panel was significantly higher than pAUC of both purely clinical and laboratory panels ($P = 0.0004$ and 0.05, respectively.)

Discussion

In this prospective study, including 141 patients from two separated cohorts of patients presenting aSAH, we have

demonstrated for the first time that elevated serum concentration of H-FABP at hospital admission was able to predict unfavorable clinical prognosis at 6 months. More importantly, the development of a multiparameter panel strategy, using a combination of blood-borne biomarkers together with a clinical score (WFNS), considerably improved unfavorable outcome prediction compared to solely clinical parameters, alone or in combination, by allowing identification of poor neurological outcome in patients with a sensitivity around 70% and a specificity of 100% following aSAH.

Identification of predictors is an important aspect of the management and study of patients with aSAH. Several clinical factors have been identified as independent predictors of patient outcome following aSAH [3, 29, 30]. Among them, clinical scores describing the patient's neurological condition at hospital admission were reported to correlate with long-term outcome [31, 32]. In line with these observations, we also showed that, when tested individually, the WFNS score at hospital admission appeared to be the strongest predictor of neurological outcome in our patient sample. In agreement with previous studies [17, 33], we found a significant correlation between the amount of blood observed in the initial CT scan and long-term neurological outcome. Although the majority of earlier studies designed age as a major independent prognostic factor [34, 35], a recent, prospectively conducted trial including 177 poor WFNS grade patients with aSAH did not find a significant association between age and outcome [36]. In our cohort, the age of patients presenting a poor GOS at 6 months was slightly but significantly higher than those with a favorable course, suggesting a potential influence of age as a prognostic factor. In contrast, we found no significant association between occurrence of vasospasm and seizure activity during hospital stay and long-term GOS outcome. Also, neither the aneurysm location site nor the treatment modality (i.e., clipping versus coiling) showed significant association with patient prognosis.

To our knowledge, this is the first study investigating the role of the recently identified, brain-related biomarkers H-FABP, NDKA and UFD-1 in the context of aSAH. These molecules have recently been shown to be reliable early blood biomarkers in ischemic stroke. H-FABP is a well-known marker for myocardial injury [37, 38] and also appears to be a potential biomarker of stroke [10, 39]. Results of the present study revealed that H-FABP was one of the best outcome predictors at 6 months (42.5% SE and 92% SP), and its performance was as high as WFNS. In addition, H-FABP was an important parameter of the panel, since its absence induces a decrease of the sensitivity from 70 to 47%.

NDKA (also called NM23-H1) is an ubiquitous enzyme that catalyzes the transfer of the terminal phosphate of ATP to (deoxy)nucleotide triphosphates via the formation of a high-energy phosphorylated intermediate. Specific expression pattern and enzymatic activity of this protein have been demonstrated in the brain [40]. In stroke, NDKA was described as an early biomarker since its level was already elevated in blood of patients within 3 h after the stroke onset [11]. In the present aSAH study, NDKA alone appeared to be an unsatisfactory predictor of outcome at 6 months. However, in combination with other parameters, its presence drastically increased the sensitivity of the panel, suggesting that its strength resides in the detection of patients not included by other predictors.

Several studies highlighted an increasing interest in S100β, a calcium-binding protein, in various brain damage disorders, and especially in aSAH [6, 41]. In these studies, elevated levels of S100β correlated with neurological deficit and outcome at 6 months or 1 year [18–20]. Our present results are in line with these observations, showing a 35% SE and 96% SP of this protein in the prediction of neurological outcome at 6 months.

The commonly used cardiac biomarker troponin I, also known as cardiac isoform of troponin I (cTnI), has previously been reported to be correlated with neurological outcome following aSAH [42]. In fact, cardiopulmonary dysfunctions could occur after aSAH, but their impacts in the mortality rate or outcome remain controversial [43, 44]. In our study, troponin I permitted to discriminate patients according to their outcome with 30% SE and 93% SP, and, in combination with other markers, it increased the sensitivity of the panel from 42 to 70%.

Outcome prediction based on a single measured biomarker or clinical score has led so far to unsatisfactory levels of sensitivity and, more importantly, specificity. Therefore, there is an urgent need to combine multiple parameters to achieve higher sensitivity without sacrificing specificity. Many studies have evaluated a multitude of classification approaches to improve the prediction performance [45]. In the present study, we used a multiparametric combination of blood-borne protein values and clinical scores. The iterative permutation-response highlighted that a six-parameter panel comprising WFNS, H-FABP, S100β, troponin I, NDKA and UFD-1 could be used for the prediction of aSAH outcome at 6 months. The six-parameter panel provided increasing prognosis sensitivity (70%) for 100% SP compared with any other parameter individually or purely clinical and laboratory panels (22% and 50% SE, respectively), when at least three out of the six predictors are above their cutoff values.

The future challenge for these biomarkers and panel is their translation in clinical practice. Several drawbacks must be solved to consider their real prospective impact in the management of SAH patients. Among them, the development of multiplex point-of-care systems should considerably reduce the time of analyses (between 15 and 30 min), making possible their use in routine clinical practice. Alternatively, new emerging ELISA technologies, such as bead suspension arrays, can also quantitate simultaneously several biomarkers in a unique patient

sample, restricting the volume need for analyses, and this in a fast and reproducible manner. Finally, the panel interpretation (binary response: positive or negative) is simple enough to be used in clinical practice.

Acknowledgments The authors thank the chief nurses and nurses of the Pitié-Salpêtrière Hospital for their remarkable work in the collection of samples. The collection of samples was funded by the Direction for Clinical Research of the Assistance Publique-Hôpitaux de Paris. This work was also kindly supported by Proteome Sciences plc.

Conflicts of interest statement Neither financial interest nor conflicts of interest are related to this publication.

References

1. Mocco J, Ransom ER, Komotar RJ, Schmidt JM, Sciacca RR, Mayer SA, Connolly ES Jr (2006) Preoperative prediction of long-term outcome in poor-grade aneurysmal subarachnoid hemorrhage. Neurosurgery 59:529–538
2. Cesarini KG, Hardemark HG, Persson L (1999) Improved survival after aneurysmal subarachnoid hemorrhage: review of case management during a 12-year period. J Neurosurg 90:664–672
3. Saveland H, Brandt L (1994) Which are the major determinants for outcome in aneurysmal subarachnoid hemorrhage? A prospective total management study from a strictly unselected series. Acta Neurol Scand 90:245–250
4. Nylen K, Csajbok LZ, Ost M, Rashid A, Karlsson JE, Blennow K, Nellgard B, Rosengren L (2006) CSF-neurofilament correlates with outcome after aneurysmal subarachnoid hemorrhage. Neurosci Lett 404:132–136
5. Nylen K, Csajbok LZ, Ost M, Rashid A, Blennow K, Nellgard B, Rosengren L (2007) Serum glial fibrillary acidic protein is related to focal brain injury and outcome after aneurysmal subarachnoid hemorrhage. Stroke 38:1489–1494
6. Stranjalis G, Korfias S, Psachoulia C, Kouyialis A, Sakas DE, Mendelow AD (2007) The prognostic value of serum S-100B protein in spontaneous subarachnoid haemorrhage. Acta Neurochir (Wien) 149:231–237, discussion 237–238
7. Oertel M, Schumacher U, McArthur DL, Kastner S, Boker DK (2006) S-100B and NSE: markers of initial impact of subarachnoid haemorrhage and their relation to vasospasm and outcome. J Clin Neurosci 13:834–840
8. Lescuyer P, Allard L, Zimmermann-Ivol CG, Burgess JA, Hughes-Frutiger S, Burkhard PR, Sanchez JC, Hochstrasser DF (2004) Identification of post-mortem cerebrospinal fluid proteins as potential biomarkers of ischemia and neurodegeneration. Proteomics 4:2234–2241
9. Burgess JA, Lescuyer P, Hainard A, Burkhard PR, Turck N, Michel P, Rossier JS, Reymond F, Hochstrasser DF, Sanchez JC (2006) Identification of brain cell death associated proteins in human post-mortem cerebrospinal fluid. J Proteome Res 5:1674–1681
10. Zimmermann-Ivol CG, Burkhard PR, Le Floch-Rohr J, Allard L, Hochstrasser DF, Sanchez JC (2004) Fatty acid binding protein as a serum marker for the early diagnosis of stroke: a pilot study. Mol Cell Proteomics 3:66–72
11. Allard L, Burkhard PR, Lescuyer P, Burgess JA, Walter N, Hochstrasser DF, Sanchez JC (2005) PARK7 and nucleoside diphosphate kinase A as plasma markers for the early diagnosis of stroke. Clin Chem 51:2043–2051
12. Allard L, Turck N, Burkhard PR, Walter N, Rosell A, Gex-Fabry M, Hochstrasser DF, Montaner J, Sanchez JC (2007) UFD1 as a blood marker for the early diagnosis of ischemic stroke. Biomarker Insights 2:155–164
13. Turck N, Vutskits L, Sanchez-Pena P, Robin X, Hainard A, Gex-Fabry M, Fouda C, Mueller M, Puybasset L, Sanchez JC (2009) A multiparameter panel method for outcome prediction following aneurysmal subarachoid hemorrhage. In: 5th International conference on biochemical markers for brain damage, 11–14 May, Lund, Sweden
14. Turck N, Vutskits L, Sanchez-Pena P, Robin X, Hainard A, Gex-Fabry M, Fouda C, Mueller M, Puybasset L, Sanchez JC (2008) From the proteomics to the clinical study: identification of a multiparameter panel for outcome prediction following aneurysmal subarachnoid hemorrhage. In: 8th Siena meeting from genome to proteome: integration and plateform completion: August 31st–September 4th, Siena, Italy
15. Turck N, Vutskits L, Sanchez-Pena P, Robin X, Hainard A, Gex-Fabry M, Fouda C, Mueller M, Puybasset L, Sanchez JC (2008) A multiparameter panel method for outcome prediction following aneurysmal subarachoid hemorrhage. In: XVII European stroke conference, 13–16 May, Nice, France
16. Drake C (1988) Report of World Federation of Neurological Surgeons Committee on a universal subarachnoid hemorrhage grading scale. J Neurosurg 68:985–986
17. Fisher CM, Kistler JP, Davis JM (1980) Relation of cerebral vasospasm to subarachnoid hemorrhage visualized by computerized tomographic scanning. Neurosurgery 6:1–9
18. Weiss N, Sanchez-Pena P, Roche S, Beaudeux JL, Colonne C, Coriat P, Puybasset L (2006) Prognosis value of plasma S100B protein levels after subarachnoid aneurysmal hemorrhage. Anesthesiology 104:658–666
19. Pereira AR, Sanchez-Pena P, Biondi A, Sourour N, Boch AL, Colonne C, Lejean L, Abdennour L, Puybasset L (2007) Predictors of 1-year outcome after coiling for poor-grade subarachnoid aneurysmal hemorrhage. Neurocrit Care 7:18–26
20. Sanchez-Pena P, Pereira AR, Sourour NA, Biondi A, Lejean L, Colonne C, Boch AL, Al Hawari M, Abdennour L, Puybasset L (2008) S100B as an additional prognostic marker in subarachnoid aneurysmal hemorrhage. Crit Care Med 36:2267–2273
21. Jennett B, Bond M (1975) Assessment of outcome after severe brain damage. Lancet 1:480–484
22. Beaudeux JL, Leger P, Dequen L, Gandjbakhch I, Coriat P, Foglietti MJ (2000) Influence of hemolysis on the measurement of S-100beta protein and neuron-specific enolase plasma concentrations during coronary artery bypass grafting. Clin Chem 46:989–990
23. McClish DK (1989) Analyzing a portion of the ROC curve. Med Decis Making 9:190–195
24. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874
25. Carpenter J, Bithell J (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 19:1141–1164

26. Hanley JA, McNeil BJ (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839–843
27. Reynolds MA, Kirchick HJ, Dahlen JR, Anderberg JM, McPherson PH, Nakamura KK, Laskowitz DT, Valkirs GE, Buechler KF (2003) Early biomarkers of stroke. Clin Chem 49:1733–1739
28. Hainard A, Tiberti N, Robin X, Lejon V, Ngoyi DM, Matovu E, Enyaru JC, Fouda C, Ndung'u JM, Lisacek F, Muller M, Turck N, Sanchez JC (2009) A Combined CXCL10, CXCL8 and H-FABP panel for the staging of human African Trypanosomiasis patients. PLoS Negl Trop Dis 3:e459
29. Lagares A, Gomez PA, Lobato RD, Alen JF, Alday R, Campollo J (2001) Prognostic factors on hospital admission after spontaneous subarachnoid haemorrhage. Acta Neurochir (Wien) 143:665–672
30. Yoshikai S, Nagata S, Ohara S, Yuhi F, Sakata S, Matsuno H (1996) A retrospective analysis of the outcomes of patients with aneurysmal subarachnoid hemorrhages: a focus on the prognostic factors. No Shinkei Geka 24:733–738
31. Gerber CJ, Lang DA, Neil-Dwyer G, Smith PW (1993) A simple scoring system for accurate prediction of outcome within four days of a subarachnoid haemorrhage. Acta Neurochir (Wien) 122:11–22
32. Hirai S, Ono J, Yamaura A (1996) Clinical grading and outcome after early surgery in aneurysmal subarachnoid hemorrhage. Neurosurgery 39:441–446 discussion 446–447
33. Brouwers PJ, Dippel DW, Vermeulen M, Lindsay KW, Hasan D, van Gijn J (1993) Amount of blood on computed tomography as an independent predictor after aneurysm rupture. Stroke 24:809–814
34. Deruty R, Pelissou-Guyotat I, Mottolese C, Amat D, Bognar L (1995) Level of consciousness and age as prognostic factors in aneurysmal SAH. Acta Neurochir (Wien) 132:1–8
35. Lagares A, Gomez PA, Alen JF, Lobato RD, Rivas JJ, Alday R, Campollo J, de la Camara AG (2005) A comparison of different grading scales for predicting outcome after subarachnoid haemorrhage. Acta Neurochir (Wien) 147:5–16 discussion 16
36. Laidlaw JD, Siu KH (2003) Poor-grade aneurysmal subarachnoid hemorrhage: outcome after treatment with urgent surgery. Neurosurgery 53:1275–1280 discussion 1280–1282
37. Suzuki M, Hori S, Noma S, Kobayashi K (2005) Prognostic value of a qualitative test for heart-type fatty acid-binding protein in patients with acute coronary syndrome. Int Heart J 46:601–606
38. Tanaka T, Sohmiya K, Kitaura Y, Takeshita H, Morita H, Ohkaru Y, Asayama K, Kimura H (2006) Clinical evaluation of point-of-care-testing of heart-type fatty acid-binding protein (H-FABP) for the diagnosis of acute myocardial infarction. J Immunoassay Immunochem 27:225–238
39. Wunderlich MT, Hanhoff T, Goertler M, Spener F, Glatz JF, Wallesch CW, Pelsers MM (2005) Release of brain-type and heart-type fatty acid-binding proteins in serum after acute ischaemic stroke. J Neurol 252:718–724
40. Dabernat S, Larou M, Masse K, Hokfelt T, Mayer G, Daniel JY, Landry M (1999) Cloning of a second nm23–M1 cDNA: expression in the central nervous system of adult mouse and comparison with nm23–M2 mRNA distribution. Brain Res Mol Brain Res 63:351–365
41. Vos PE, van Gils M, Beems T, Zimmerman C, Verbeek MM (2006) Increased GFAP and S100beta but not NSE serum levels after subarachnoid haemorrhage are associated with clinical severity. Eur J Neurol 13:632–638
42. Yarlagadda S, Rajendran P, Miss JC, Banki NM, Kopelnik A, Wu AH, Ko N, Gelb AW, Lawton MT, Smith WS, Young WL, Zaroff JG (2006) Cardiovascular predictors of in-patient mortality after subarachnoid hemorrhage. Neurocrit Care 5:102–107
43. Schuiling WJ, Dennesen PJ, Tans JT, Kingma LM, Algra A, Rinkel GJ (2005) Troponin I in predicting cardiac or pulmonary complications and outcome in subarachnoid haemorrhage. J Neurol Neurosurg Psychiatry 76:1565–1569
44. Ramappa P, Thatai D, Coplin W, Gellman S, Carhuapoma JR, Quah R, Atkinson B, Marsh JD (2008) Cardiac troponin-I: a predictor of prognosis in subarachnoid hemorrhage. Neurocrit Care 8:398–403
45. Germanson TP, Lanzino G, Kongable GL, Torner JC, Kassell NF (1998) Risk classification after aneurysmal subarachnoid hemorrhage. Surg Neurol 49:155–163

6

A Combined CXCL10, CXCL8 and H-FABP Panel for the Staging of Human African Trypanosomiasis Patients

This chapter describes a second application of PanelomiX, on Human African trypanosomiasis (HAT), also known as sleeping sickness. It is a parasitic tropical disease that progresses from the first, haemolymphatic stage to a neurological second stage where the parasites invades the central nervous system. As treatment depends on the stage of disease, there is a critical need for tools that efficiently discriminate the two stages of HAT. At the same time, and because the parasite hits essentially poor people in developing African countries, an additional requirement for a potential panel is an extremely low cost. Therefore, only a very limited number of biomarkers can be included in the combination to indicate the CNS invasion by the parasite.

One hundred Cerebrospinal fluid (CSF) samples originating from parasitologically confirmed Trypanosoma brucei gambiense patients were analysed: 21 from stage 1 (no trypanosomes in CSF and ≤ 5 WBC/mL) and 79 from stage 2 (trypanosomes in CSF or > 5 WBC/mL) patients. The concentration of H-FABP, GSTP-1 and S100b in CSF was measured by ELISA. The levels of thirteen inflammation-related proteins (IL-1ra, IL-1b, IL-6, IL-9, IL-10, G-CSF, VEGF, IFN-c, TNF-a, CCL2, CCL4, CXCL8 and CXCL10) were determined by bead suspension arrays. Patients were staged on the basis of CSF white blood cell (WBC) count and presence of parasites in CSF.

CXCL10 most accurately distinguished stage 1 and stage 2 patients, with a sensitivity of 84% and specificity of 100%. A panel of 3 proteins, CXCL10, CXCL8 and H-FABP, improved the detection of stage 2 patients to 97% sensitivity and 100% specificity.

This study highlights the value of CXCL10 as a single biomarker for staging T. b. gambiense-infected HAT patients. Further combination of CXCL10 with H-FABP and CXCL8 results in a panel that efficiently rules in stage 2 HAT patients. As these molecules could potentially be markers of other CNS infections and disorders, these results should be validated in a larger multi-centric cohort including other inflammatory diseases such as cerebral malaria and active tuberculosis.

This article is mainly the work of the two first authors. My contribution is centered around the data analysis. I determined the panel and conducted the ROC analysis.

OPEN ACCESS Freely available online

PLoS NEGLECTED TROPICAL DISEASES

A Combined CXCL10, CXCL8 and H-FABP Panel for the Staging of Human African Trypanosomiasis Patients

Alexandre Hainard[1], Natalia Tiberti[1], Xavier Robin[1], Veerle Lejon[2], Dieudonné Mumba Ngoyi[3], Enock Matovu[4], John Charles Enyaru[5], Catherine Fouda[1], Joseph Mathu Ndung'u[6], Frédérique Lisacek[7], Markus Müller[7], Natacha Turck[1], Jean-Charles Sanchez[1]*

1 Biomedical Proteomics Research Group, Medical University Centre, Geneva, Switzerland, **2** Department of Parasitology, Institute of Tropical Medicine, Antwerp, Belgium, **3** Institut National de Recherche Biomedicale, Kinshasa, D.R. Congo, **4** Department of Veterinary Parasitology and Microbiology, Faculty of Science, Makerere University, Kampala, Uganda, **5** Department of Biochemistry, Faculty of Science, Makerere University, Kampala, Uganda, **6** Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland, **7** Swiss Institute of Bioinformatics, Medical University Centre, Geneva, Switzerland

Abstract

Background: Human African trypanosomiasis (HAT), also known as sleeping sickness, is a parasitic tropical disease. It progresses from the first, haemolymphatic stage to a neurological second stage due to invasion of parasites into the central nervous system (CNS). As treatment depends on the stage of disease, there is a critical need for tools that efficiently discriminate the two stages of HAT. We hypothesized that markers of brain damage discovered by proteomic strategies and inflammation-related proteins could individually or in combination indicate the CNS invasion by the parasite.

Methods: Cerebrospinal fluid (CSF) originated from parasitologically confirmed *Trypanosoma brucei gambiense* patients. Patients were staged on the basis of CSF white blood cell (WBC) count and presence of parasites in CSF. One hundred samples were analysed: 21 from stage 1 (no trypanosomes in CSF and $\leq$5 WBC/μL) and 79 from stage 2 (trypanosomes in CSF and/or $>$5 WBC/μL) patients. The concentration of H-FABP, GSTP-1 and S100β in CSF was measured by ELISA. The levels of thirteen inflammation-related proteins (IL-1ra, IL-1β, IL-6, IL-9, IL-10, G-CSF, VEGF, IFN-γ, TNF-α, CCL2, CCL4, CXCL8 and CXCL10) were determined by bead suspension arrays.

Results: CXCL10 most accurately distinguished stage 1 and stage 2 patients, with a sensitivity of 84% and specificity of 100%. Rule Induction Like (RIL) analysis defined a panel characterized by CXCL10, CXCL8 and H-FABP that improved the detection of stage 2 patients to 97% sensitivity and 100% specificity.

Conclusion: This study highlights the value of CXCL10 as a single biomarker for staging *T. b. gambiense*-infected HAT patients. Further combination of CXCL10 with H-FABP and CXCL8 results in a panel that efficiently rules in stage 2 HAT patients. As these molecules could potentially be markers of other CNS infections and disorders, these results should be validated in a larger multi-centric cohort including other inflammatory diseases such as cerebral malaria and active tuberculosis.

Citation: Hainard A, Tiberti N, Robin X, Lejon V, Ngoyi DM, et al. (2009) A Combined CXCL10, CXCL8 and H-FABP Panel for the Staging of Human African Trypanosomiasis Patients. PLoS Negl Trop Dis 3(6): e459. doi:10.1371/journal.pntd.0000459

Editor: Jayne Raper, New York University School of Medicine, United States of America

Received January 29, 2009; **Accepted** May 15, 2009; **Published** June 16, 2009

Funding: This work is supported by the Foundation for Innovative New Diagnostics (FIND). The THARSAT study, including a PhD grant of DMN, received financial support from the Belgian Ministry of Foreign Affairs, Directorate General for Development Co-operation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Jean-Charles.Sanchez@unige.ch

Introduction

Human African trypanosomiasis (HAT), also called sleeping sickness, is a parasitic disease that occurs in sub-Saharan Africa. More than sixty million people are at risk of being infected. The World Health Organization (WHO) has reported impressive progress since 1995 in the control of HAT, leading to a substantial reduction of new cases detected yearly to 10'800 in 2007. The total number of cases is now estimated to be between 50'000 and 70'000 per year [1].

The parasite that causes HAT belongs to the *Trypanosoma brucei* family with two subspecies, *Trypanosoma brucei gambiense* and *Trypanosoma brucei rhodesiense*, responsible for the human disease. Trypanosomes are transmitted to humans by the bite of a tsetse fly and are initially confined to the blood, lymph nodes and peripheral tissues. This corresponds to the first stage (early stage; or haemolymphatic stage) of the disease. After an unknown period that varies from weeks to months, the parasites invade the central nervous system (CNS). This is called the second stage (late stage; or neurologic; or meningo-encephalitic stage) of HAT.

Clinical symptoms of HAT are not specific for the disease, and definite diagnosis is always based on parasitological examination of body fluids. The card agglutination test for trypanosomiasis (CATT), an assay that is based on trypanosome-specific antibody detection, is widely used for mass screening. However, it suffers from limited sensitivity and restricted to the *T. b. gambiense* form of

Author Summary

The actual serological and parasitological tests used for the diagnosis of human African trypanosomiasis (HAT), also known as sleeping sickness, are not sensitive and specific enough. The card agglutination test for trypanosomiasis (CATT) assay, widely used for the diagnosis, is restricted to the *gambiense* form of the disease, and parasitological detection in the blood and cerebrospinal fluid (CSF) is often very difficult. Another very important problem is the difficulty of staging the disease, a crucial step in the decision of the treatment to be given. While eflornithine is difficult to administer, melarsoprol is highly toxic with incidences of reactive encephalopathy as high as 20%. Staging, which could be diagnosed as early (stage 1) or late (stage 2), relies on the examination of CSF for the presence of parasite and/or white blood cell (WBC) counting. However, the parasite is rarely found in CSF and WBC count is not standardised (cutoff set between 5 and 20 WBC per μL). In the present study, we hypothesized that an early detection of stage 2 patients with one or several proteins in association with clinical evaluation and WBC count would improve staging accuracy and allow more appropriate therapeutic interventions.

the disease [2]. A positive parasitological diagnosis must always be followed by stage determination, which is performed by examination of the cerebrospinal fluid (CSF). This is a vital step in the diagnostic process, as the treatment differs depending on the stage of the disease. If HAT patients are not treated, they always die [3–5]. Early stage drugs are inefficient for late stage patients, and additionally, melarsoprol (MelB or Arsobal), which has been the most widely used drug to treat late stage patient, has itself an overall mortality rate of 5% due to its toxicity [6]. As a consequence, melarsoprol has in many countries been replaced by eflornithine as the first line treatment for *T. b. gambiense* infections but the latter drug suffers from important logistic constraints.

WHO defined late-stage HAT by the following criteria: presence of trypanosomes in CSF and/or an elevated WBC count above 5/μL of CSF [7]. However, presence of WBC in the CSF is not specific for the disease and parasite detection methods are not sensitive enough [8]. Furthermore, recent studies suggest the need to increase the cutoff between the first and second stages to 10 or 20 WBC/μL [2,8,9]. This has contributed to the concept of a potential intermediate stage of HAT with CSF WBC count >5 and ≤ 20 WBC/μL [10]. There is therefore a critical need for a reliable and efficient staging tool that would replace or complement trypanosome detection and WBC count.

Parasite migration and invasion of the CNS causes a neuroinflammatory process, associated with activation of microglial cells and astrocytes [11,12], and infiltration of the CNS with leukocytes (predominantly mononuclear cells) [13]. Cytokines and chemokines are known to be actively involved in this process. Thus, TNF-α, IL-6, CXCL8 and IL-10 concentrations have been demonstrated to be elevated in the CSF of late-stage patients [11,14] and the IFN-γ level has been reported as associated with the severity of the late stage disease [15]. The levels of CCL2, IL-1β and CXCL8 have also been correlated with presence of parasites in the CSF and neurological signs in HAT patients [16]. Additionally, levels of IL-1ra, G-CSF, VEGF, CCL4 and CXCL10 were found modulated in either the CSF or plasma of patients suffering from cerebral malaria [17–19], and could potentially be also modulated in HAT patients.

Proteomic analysis of human body fluids has become an important approach for biomarkers discovery [20]. In this context, we recently explored the concept of *post-mortem* CSF as a model of massive and global brain insult [21], which allowed the identification of potential brain damage biomarkers by proteomics strategies. Indeed, heart-fatty acid binding protein (H-FABP), identified from *post-mortem* CSF, has been validated as a marker of stroke [22] and Creutzfeldt-Jakob disease [23], respectively. Similarly, GSTP-1 was also found over-expressed in *post-mortem* CSF [24] compared to *ante-mortem*, and was recently validated as an early diagnostic marker of stroke and traumatic brain injury (Turck *et al.* Personal communication). Additionally, S100β protein has already been demonstrated to be a marker of blood-brain barrier (BBB) and neuronal damage [25] as well as a useful serum biomarker of CNS injury and a potential tool for predicting clinical outcome after brain damage [26].

In this context, we hypothesized that markers of brain damage discovered by proteomic strategies as well as inflammation-related proteins could individually or in combination indicate the CNS invasion by the trypanosome parasite. We measured the CSF concentrations of H-FABP, GSTP-1, S100β and thirteen inflammation-related proteins (IL-1ra, IL-1β, IL-6, IL-9, IL-10, G-CSF, VEGF, IFN-γ, TNF-α, CCL2, CCL4, CXCL8 and CXCL10) and evaluated their potential for staging the disease.

Material and Methods

Samples

Samples originated from a prospective observational study on shortening of post treatment follow-up in *gambiense* human African trypanosomiasis (THARSAT), conducted between 2005 and 2008 at Dipumba hospital in Mbuji-Mayi (Kasai Oriental province, Democratic Republic of the Congo). Details of the THARSAT study design and results are reported elsewhere (D. Mumba Ngoyi, in preparation). The study protocol was approved by the Ministry of Health, Kinshasa, DRC and by the Ethical Committee of the University of Antwerp, Belgium. Briefly, 360 *T. b. gambiense* patients in total were enrolled into the THARSAT study. Inclusion criteria were 1° confirmed presence of trypanosomes in lymph nodes, blood or CSF; 2°≥12 years old and; 3° living within a perimeter of 100 km around Mbuji-Mayi. Exclusion criteria were 1° pregnancy; 2° no guarantee for follow-up; 3° moribund; 4° haemorrhagic CSF before treatment and; 5° presence of another serious illness (active tuberculosis - treated or not, bacterial or cryptococcal meningitis). HIV and malaria were not considered as exclusion criteria. Each patient underwent a clinical examination. Staging of disease was based on CSF examination. WBC count was performed in disposable cell counting chambers (Uriglass, Menarini) and was performed in duplicate when the first count was <20 cells/μL. Trypanosomes were searched for in CSF by direct examination prior or during the cell counting procedure, followed by the modified single centrifugation method [27]. Second stage patients were defined as having >5 WBC/μL and/or trypanosomes in the CSF. First stage patients were defined as having 0–5 WBC/μL and no trypanosomes in the CSF. Patients having >5 and ≤ 20 WBC/μL and no trypanosomes in CSF were defined and treated as stage 2 patients, but highlighted as being in the potential intermediate stage. Patients or their responsible were informed about the study objectives and modalities and were asked to provide written consent. Treatment was provided according to the guidelines of the national control program for HAT (PNLTHA).

CSF samples were centrifuged immediately after collection. The supernatant remaining after the diagnostic procedure was

aliquoted, stored and shipped frozen at −20°C or below. For the study reported here, a total of 100 CSF samples, taken before treatment, were tested. These samples originated from 21 stage 1 (S1) and 79 stage 2 patients (S2). S1 patients were age and sex matched with 21 S2 patients. Remainder S2 patients were chosen in order to obtain homogenous median age values. Patients were classified into three categories of neurological signs; absent (no neurological signs), moderate (at least one major neurological sign but no generalised tremors) or severe (at least two major neurological signs including generalised tremors). Major neurological signs were defined as: daytime somnolence, sensory and gait disturbances, presence of primitive reflexes (Babinski's sign, palmo-mental reflex, perioral reflex), modified tendon reflexes (exaggeration or abolition), abnormal movements such as tremor (fine, diffuse and generalised). Neurological signs were not reported for two patients.

S100β, H-FABP and GSTP-1 measurements

The concentration of S100β was measured using a commercially available sandwich ELISA assay kit (Abnova, Taiwan) following the manufacturer's instructions. Briefly, calibrators, Quality control (QC) and CSF samples diluted 1:4 were incubated 2 hours on microtiter strips pre-coated with polyclonal anti-cow S100β antibodies. After 3 washes, horseradish peroxidase (HRP) labelled anti-human S100β antibodies were added, incubated for 90 minutes and washed again before addition of the substrate solution (tetramethylbenzidine). Color development was stopped with sulphuric acid and absorbance was read on a Vmax Kinetic microplate reader, (Molecular Devices Corporation, Sunnyvale, CA, U.S.A.) at a wavelength of 450 nm.

H-FABP concentration was also determined using a commercially available ELISA kit (Hycult Biotechnology, Uden, Netherlands) according to the manufacturer's instructions. CSF samples (non-diluted) and standards were incubated (1 hour) together with peroxidase conjugated secondary antibodies in microtiter wells coated with antibodies recognizing human H-FABP. After 3 washes, tetramethylbenzidine was added and color development was stopped by adding citric acid.

The concentration of GSTP-1 was determined using a homemade ELISA as described by Allard *et al.* [28]. Briefly, biotinylated anti-GSTP-1 antibodies (2 µg/mL) (Biosite, California, USA) were coated onto a 96-well Reacti-Bind NeutrAvidin coated Black Plates (Pierce, Rockford, IL) for 1 hour at 37°C. After 3 washes, CSF samples (diluted 1:4), quality controls and standards (recombinant GSTP-1 at concentrations ranging from 0 to 100 ng/mL) were incubated for 1 hour at 37°C, and followed by a washing step. Alkaline phosphatase conjugated antibodies against human GSTP-1 (Biosite, California, USA) at 2 µg/mL were added and incubated for 1 hour at 37°C. After 3 washes, Attophos AP fluorescent substrate (Promega, Madison, WI) was added and plates were read immediately on a SpectraMax GEMINI-XS (Molecular Devices Corporation, Sunnyvale, CA, U.S.A.) plate reader, using the kinetic mode. Vmax values were automatically calculated by the instruments based on relative fluorescence units (RFU) ($\lambda_{excitation} = 444$ nm and $\lambda_{emission} = 555$ nm).

Concentrations of S100β, H-FABP and GSTP-1 in the CSF samples were back-calculated using a linear calibration curve based on measured standards values.

Bead suspension array

The levels of thirteen cytokines and chemokines (IL-1ra, IL-1β, IL-6, IL-9, IL-10, G-CSF, VEGF, IFN-γ, TNF-α, CCL2, CCL4, CXCL8 and CXCL10) were determined using the Bioplex bead suspension arrays according to the manufacturer's instructions (Bio-Rad, Hercules, CA). Briefly, thirteen sets of color-coded polystyrene beads were conjugated separately with one of the thirteen different antibodies against the molecule of interest. All the sets were then mixed together by the supplier and delivered ready-to-use. An equal amount of beads was added to each well of a 96-well filter plate. After a series of washes, standards and samples (diluted 1:4) were added and incubated for 30 minutes at room temperature. After washing, a mix of the corresponding thirteen biotinylated detection antibodies was added and incubated 30 minutes at room temperature. After washing, streptavidin-phycoerythrin (streptavidin-PE) was added for 10 minutes. After a last series of washes, beads were re-suspended in the provided assay buffer and each well was aspirated using the Bio-Plex system. Each bead was identified and the corresponding target simultaneously quantified based respectively on bead color and fluorescence. The concentration of each target was automatically calculated by the Bio-Plex Manager software using corresponding standard curve (5-PL regression) obtained from recombinant protein standards.

Data and statistical analysis

Descriptive statistics were performed using the SPSS (version 16.0, SPSS Inc., Chicago, IL, USA) and GraphPad Prism (version 4.03, GraphPad software Inc., San Diego, CA, USA) software. Because none of the markers presented a normal distribution in concentrations (Kolmogorov-Smirnov test), differences between groups were tested with non-parametric Mann-Whitney U test (comparison between two groups) and Kruskal-Wallis test followed by Dunn's post-hoc test (comparison between three groups). Statistical significance for these tests was set at 0.05 (2-tailed tests). The stage, the presence of the parasite in CSF and the severity of neurological signs were successively considered as the dependent variables. The different marker concentrations were considered as independent variables. Bivariate non-parametric correlations using the Spearman correlation coefficient were carried out with statistical significance set at 0.01 (2-tailed tests).

To calculate the sensitivity and specificity of each individual predictor with respect to staging, the specific receiver operator characteristic (ROC) curve of each analyte was determined and the cutoff value was selected as the threshold predicting stage 2 patients with 100% of specificity (Figure S1).

Aabel (version 2.4.2, Gigawiz Ltd. Co., Tulsa, OK, USA) was used for box plots, SPSS for scatter plots and R (version 2.8.0) [29] was used for plotting ROC curves.

Panel development

Panel selection was mainly performed as described by Reynolds *et al.* [30]. Briefly, the optimized cutoff values were obtained by modified iterative permutation-response calculations (rule-induction-like, RIL) using only the molecules that presented a p value<0.0001 (Mann-Whitney U test), an AUC above 75% and a significant Spearman correlation with WBC above 0.4 (Table 2). Each cutoff value was changed iteratively by quantile of 2% increment and sensitivity was determined after each iteration until a maximum sensitivity was achieved for 100% specificity. The permutation–response calculations were conducted using a PERL program (ActivePerl version 5.10.0.1004, ActiveState Software Inc.) and data were coded in CSV format.

Results

Biomarker concentration as a function of disease stage

The main characteristics of the 100 patients evaluated in this study are presented in Table 1. The analytes were classified into

Table 1. Characteristics of the studied population.

		Stage 1	Stage 2
Population	n	21	79
Gender	Male	8	51
	Female	13	28
Age	Median (range)	32.0 (14–60)	33.0 (13–65)
WBC/μl	Median (range)	2 (0–5)	126 (6–6304)
Parasite in CSF	N	0	64
Neurological signs*	Absence	11	11
	Moderate	10	51
	Severe	0	15
>5 and ≤20 WBCμL No trypanosomes in CSF**	N	0	8

*Neurological signs were not reported for two patients.
**Correspond to the number of patients highlighted as being in the potential intermediate stage.
doi:10.1371/journal.pntd.0000459.t001

three groups, based on the results presented in Table 2. Criteria for the classification were the significance (Mann-Whitney U test), the AUC and the correlation with WBC. In the first group (GR1) comprising IL-1ra, G-CSF, CCL4, and VEGF, no significant difference in CSF concentrations between the two stages of HAT was observed. The second group (GR2) encompassed IFN-γ, IL-9, CCL2 and S100β, for which concentrations in the CSF were significantly different between stage 1 and stage 2 patients ($0.001<p<0.01$, Mann-Whitney U test). The third group (GR3) included GSTP-1, H-FABP, TNF-α, IL-1β, IL-6, IL-10, CXCL8 and CXCL10, for which the difference between stages was highly significant ($p<0.0001$, Mann-Whitney U test) (Figure S2).

To assess the sensitivity and specificity of these analytes for S2 HAT, ROC curves were built. GR1 and GR2 had a low to medium area under ROC curve (AUC) ranging from 54 to 70% and also displayed a low sensitivity in detecting S2 patients (4–13% for GR1 and 10–44% for GR2, see Table 2) at a predefined specificity of 100%. GR3 showed higher AUC (79–95%), and sensitivities for identification of S2 patient up to 84% (Table 2). CXCL10 appeared then as the most accurate predictor for staging, as, with a cutoff set at 2080 pg/mL, this molecule identified 66 out of 79 late stage patients and ruled-out all the early-stage patients.

Correlation between WBC and biomarker concentrations

As the white blood cell count was one of the two reference staging parameters, we investigated the correlation between the concentrations of the sixteen biomarkers and the number of WBC in CSF (Table 2). There was no significant correlation in the concentrations of the first and second group of analytes (GR1 and GR2) with WBC, except for S100β, which had a significant but low Spearman rho coefficient (0.269, $p<0.01$). Otherwise, strong correlations were observed between WBC and the concentrations of GR3 biomarkers (GSTP-1, IL-1β, IL-6, H-FABP, TNF-α, IL-10, CXCL8 and CXCL10), with Spearman rho ranging from 0.417 to 0.732 (Table 2 and Figure 1). The levels of GR3 molecules in 8 potential intermediate stage patients (parasite not detected in CSF and having >5 and ≤20 WBC/μL) demonstrated the intermediate behaviour of this category with some patients appearing as S1 and others as S2 patients (Figure 1). Based on the above results, only the GR3 molecules (GSTP-1, IL-1β, IL-6, H-FABP, TNF-α, IL-10, CXCL8 and CXCL10) were selected for further analyses.

Table 2. Detailed results for all the molecules tested in respect with the stage of the disease.

		Absence of parasite and ≤5 WBC/μl	Presence of parasite and/or >5 WBC/μl	Mann-Whitney U test	Correlation with WBC	ROC curve		
	Markers	Median (range)	Median (range)	p value	(spearman rho)	% AUC	Cutoff [pg/mL]	Sensitivity, % (95% CI)[a]
GR3	CXCL10	347.3 (24.3–2048.8)	14130.0 (24.3–128900.0)	<0.0001	0.625**	95	>2080.0	84 (74–91)
	CXCL8	56.9 (1.3–96.5)	178.9 (1.6–1791.0)	<0.0001	0.557**	94	>97.1	82 (72–90)
	IL-10	6.7 (0.9–19.6)	74.5 (2.1–573.1)	<0.0001	0.702**	89	>20.0	80 (69–88)
	TNF-α	3.3 (0.5–8.4)	22.5 (1.0–295.4)	<0.0001	0.636**	93	>8.5	78 (68–87)
	H-FABP	226.4 (19.8–564.0)	748.3 (0.0–16680.0)	<0.0001	0.417**	86	>571.8	62 (50–73)
	IL-6	5.0 (0.2–57.7)	63.8 (0.8–3286.0)	<0.0001	0.732**	94	>58.0	52 (40–63)
	IL-1β	0.1 (0.1–0.7)	0.6 (0.1–42.2)	<0.0001	0.445**	80	>0.7	48 (37–60)
	GSTP-1	1272.9 (149.7–5026.9)	3014.0 (61.2–75810.0)	<0.0001	0.437**	79	>5078.0	24 (15–35)
GR2	IFN-γ	68.7 (8.6–209.2)	100.4 (1.7–995.5)	0.0049	0.094	70	>210.9	10 (4–19)
	IL-9	23.4 (3.6–44.5)	30.7 (3.6–209.6)	0.0051	0.041	70	>45.0	23 (14–34)
	S100β	43.2 (4.9–113.0)	78.4 (0.0–353.0)	0.0053	0.269**	70	>114.3	29 (19–40)
	CCL2	428.1 (58.6–632.9)	590.2 (15.8–5391.0)	0.0055	0.156	70	>664.7	44 (33–56)
GR1	G-CSF	43.4 (2.4–209.8)	63.2 (2.0–785.9)	0.0866 (ns)	−0.029	62	>281.7	4 (1–11)
	IL-1ra	817.3 (128.6–3087.6)	782.0 (34.0–11760.0)	0.5229 (ns)	−0.065	55	>3092.0	13 (6–22)
	CCL4	94.2 (1.5–301.0)	91.9 (5.4–753.9)	0.5423 (ns)	−0.143	54	>316.6	5 (1–12)
	VEGF	48.3 (20.0–215.7)	49.4 (3.5–1009.0)	0.9393 (ns)	−0.105	54	>222.4	9 (4–17)

[a]Sensitivity was set for a specificity of 100% (95% CI, 84–100).
**Correlation is significant at the 0.01 level (2-tailed).
doi:10.1371/journal.pntd.0000459.t002

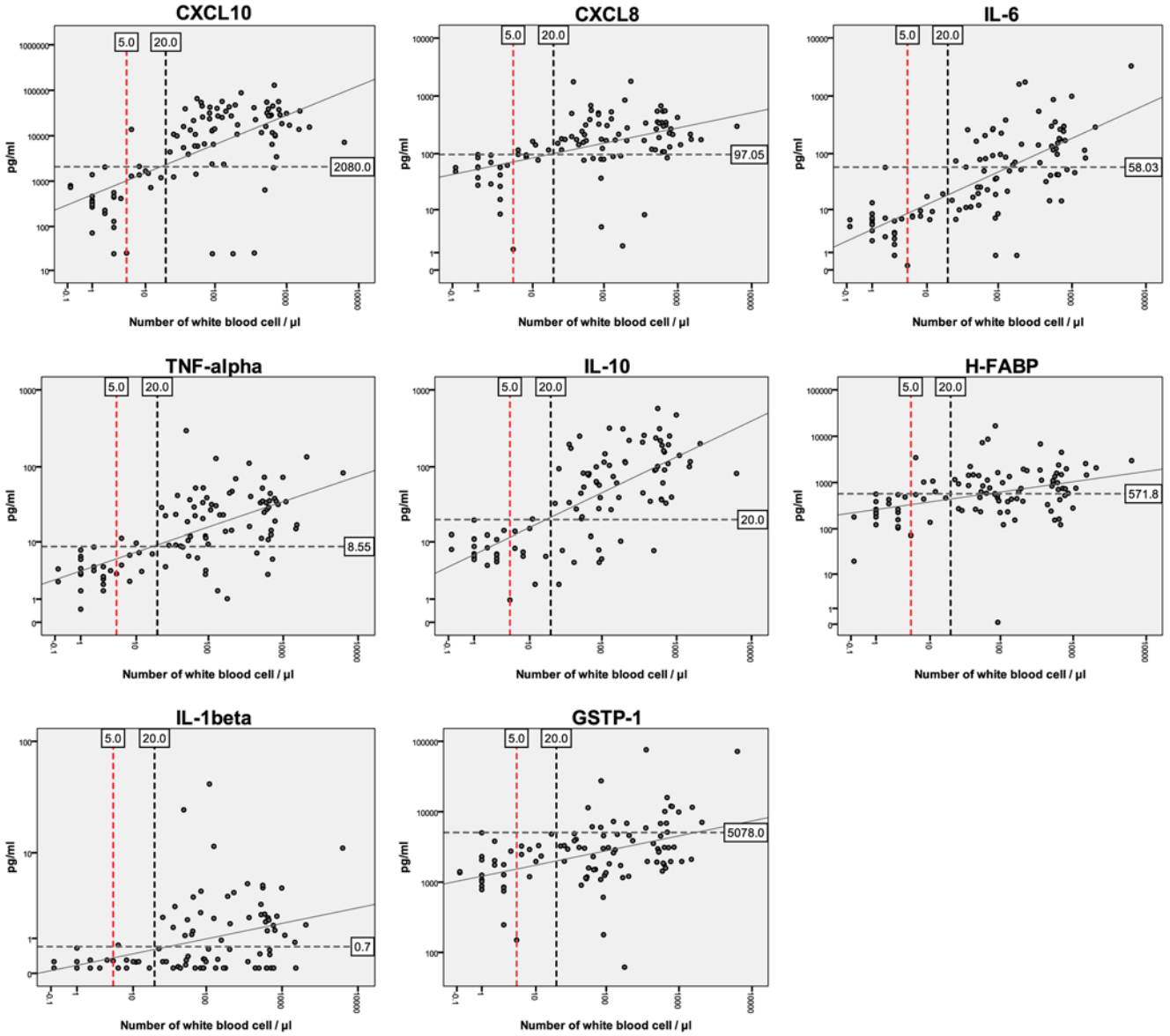


Figure 1. Scatter plots correlating the level of GR3 molecules with the WBC count. The horizontal dashed line corresponds to the cutoff value for the molecule that discriminates between S1 and S2 patients with a specificity of 100%. The left vertical dashed line corresponds to the WBC count cutoff value used for staging. The second vertical dashed line indicates the suggested cutoff value for staging. Patients between these lines (>5 and ≤20 WBC/µL) corresponded to potential intermediate stage patients. The diagonal line corresponds to the linear regression.
doi:10.1371/journal.pntd.0000459.g001

Parasites in CNS and biomarker concentrations

GR3 molecule concentrations were classified according to the absence/presence of trypanosomes in CSF. GSTP-1, IL-1β, IL-6, H-FABP, TNF-α, IL-10, CXCL8 and CXCL10 concentrations were significantly increased in patients with parasites in CSF (Figure 2 and Table S1). The six biomarkers associated with inflammation had a lower p value (<0.0001, Mann-Whitney U test) and higher AUC (ranging from 78% to 89%) than H-FABP and GSTP-1 ($0.001<p<0.05$, Mann-Whitney U test, AUCs of 69% and 64% respectively). Additionally, when only S2 patients were analysed, CXCL10, IL-10 and TNF-α levels still demonstrated a significant difference between patients with or without trypanosomes in CSF ($p<0.05$, Dunn's post-hoc test, Table S1).

Neurological signs and biomarker concentrations

The patients were classified with respect to the neurological signs reported (absence, moderate or severe) (Figure 3). All the GR3 molecules except GSTP-1 showed a significant increase in concentration associated with higher severity of neurological signs ($p<0.05$, Kruskal-Wallis test). Indeed, CXCL10, CXCL8, IL-6, IL-10, IL-1β, and TNF-α concentrations were significantly different between patients without neurological signs and severe neurological signs ($p<0.05$, Dunn's post-hoc test), as well as between patients with moderate and severe neurological signs ($p<0.05$, Dunn's post-hoc test). H-FABP level was significantly different between patients without neurological signs and severe neurological signs ($p<0.05$, Dunn's post-hoc test). Only the concentrations of CXCL10, IL-10 and TNF-α could distinguish

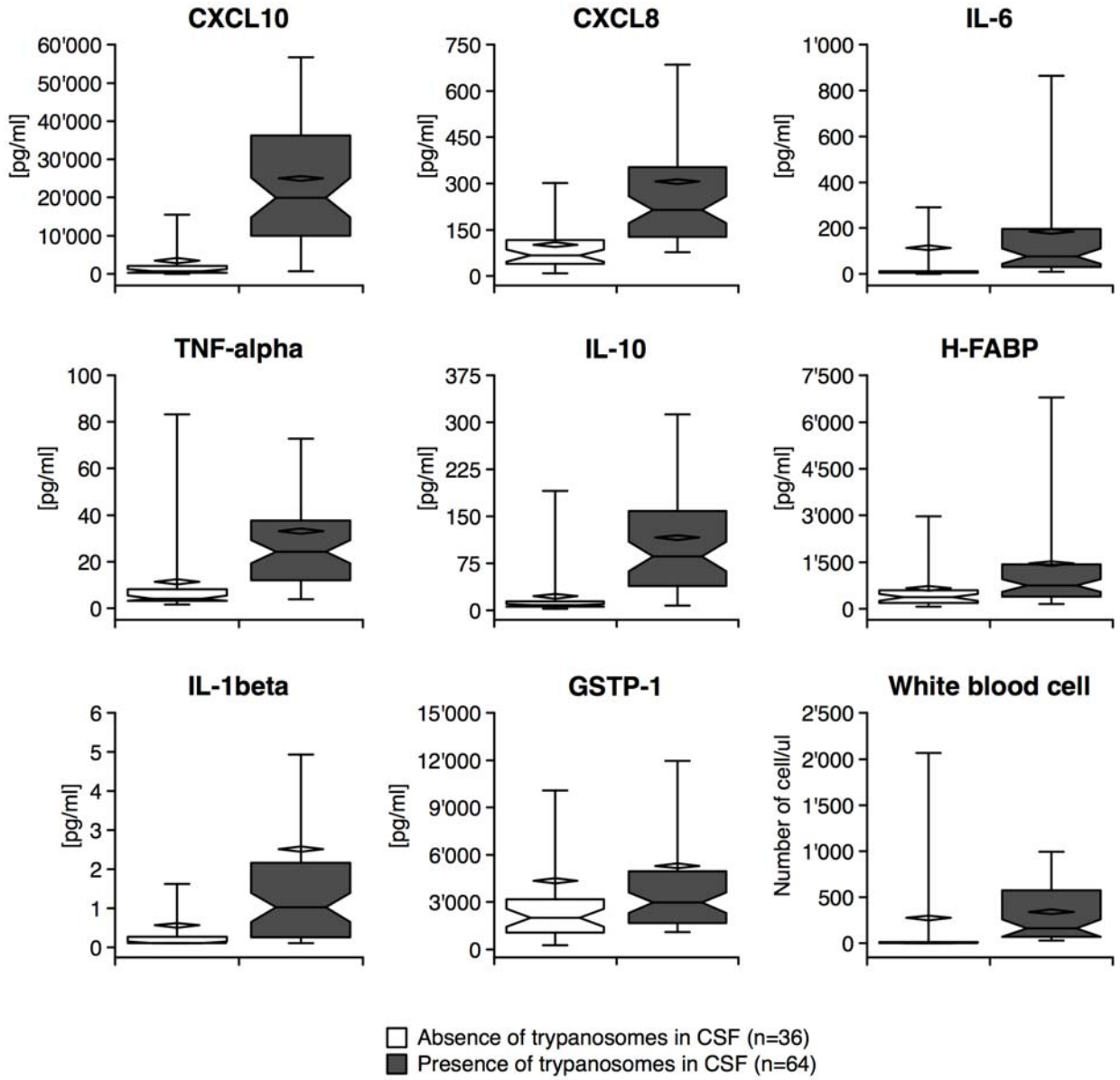


Figure 2. Box-plot of GR3 molecules and WBC classified according to the presence of the parasite in CSF. Median and mean are represented as a solid line in the box and a diamond respectively. Whisks are defined as 5^{th}–95^{th} percentile without outliers. Half-width of the notch was calculated automatically by the software.
doi:10.1371/journal.pntd.0000459.g002

between absence and moderate neurological signs ($p<0.05$, Dunn's post-hoc test).

Panel selection

In an effort to improve the global sensitivity of molecules in the prediction of second stage HAT, the GR3 molecules were combined using the rule induction like (RIL) approach. This resulted in the identification of a three-molecule panel characterized by CXCL10, CXCL8 and H-FABP (cutoff values were set at 2080.0, 97.1 and 571.8 pg/mL, respectively). A positive test (leading to identification of S2 patient) was obtained as soon as one of the three molecules included in the panel was above its cutoff value (Table 3). The panel had a sensitivity of 97% (95% CI, 91–100%) and, by definition, a specificity of 100% (95% CI, 84–100%). This means that the panel could identify 77 out of 79 stage 2 patients, and ruled-out all the 21 stage 1 patients. Out off the 77 ruled-in S2 patients, 5 were CXCL10 positive only (>2080.0 pg/mL), 6 CXCL8 positive only (>97.1 pg/mL) and 3 H-FABP positive only (>571.8 pg/mL). The rest of ruled-in S2 patients were identified with either 2 positive molecules (n = 23) or 3 positive molecules (n = 40). When this panel was applied on the intermediate stage patients (eight patients having >5 and ≤20 WBC/μL and no trypanosomes in CSF) only one patient gave a negative test response and thus 7 out of 8 patients were classified as S2.

Discussion

In this study, including early and late stage HAT patients (n = 100), we evaluated sixteen molecules as potential staging markers of HAT, to replace or complement trypanosome

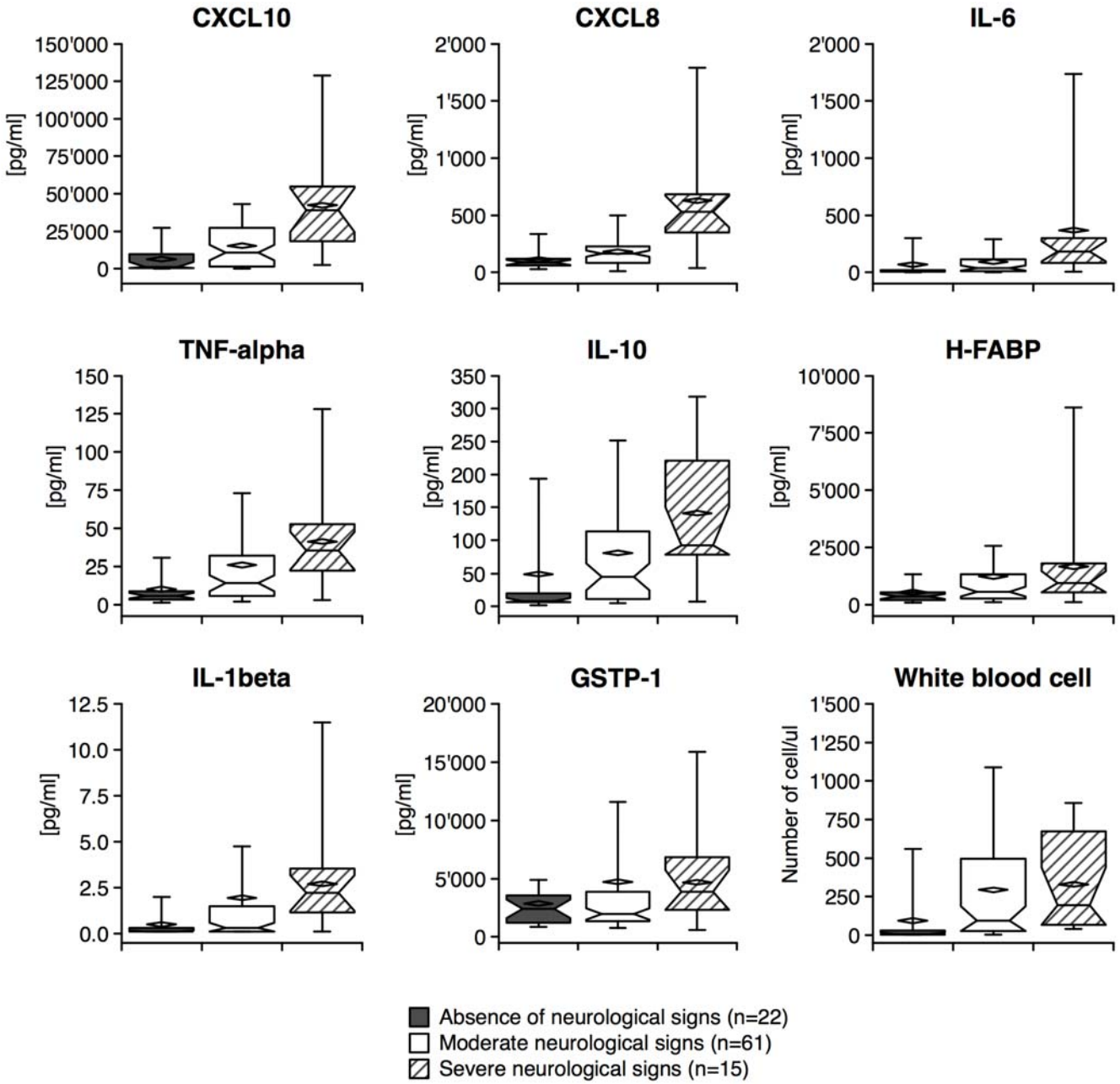


Figure 3. Box-plot of GR3 molecules and WBC classified according to the neurological signs. Median and mean are represented as a solid line in the box and a diamond respectively. Whisks are defined as 5^{th}–95^{th} percentile without outliers. Half-width of the notch was calculated automatically by the software. Neurological signs of two patients were not reported (n = 98).
doi:10.1371/journal.pntd.0000459.g003

Table 3. Detailed results for the three molecule panel in respect with the stage of the disease.

	Markers	Number of negative test	Number of positive test	Mann-Whitney U test , p value	% AUC (ROC curve)	Panel cutoff	Sensitivity, % (95% CI)[a]
Panel	CXCL10, CXCL8, H-FABP	23	77	<0.0001	99	≥1 molecule above its cutoff value[b]	97 (91–100)

[a]Sensitivity was set for a specificity of 100% (95% CI, 84–100).
[b]Cutoff values: CXCL10>2080.0 pg/mL, CXCL8>97.1 pg/mL and H-FABP>571.8 pg/mL.
doi:10.1371/journal.pntd.0000459.t003

detection and WBC count. Eight of these molecules, CXCL10, CXCL8, IL-6, IL-10, IL-1β, TNF-α, H-FABP and GSTP-1, presented concentrations significantly elevated in the CSF of late-stage HAT patients. We demonstrated that the CSF concentration of CXCL10 is highly elevated in stage 2 patients when compared to stage 1, highlighting this molecule as a potential new staging marker for sleeping sickness. A combinatorial approach has been applied in staging of HAT, in order to improve the sensitivity. This method has led to the identification of a panel consisting of CXCL10, CXCL8 and H-FABP, that identified late-stage patients with a sensitivity of 97% at 100% specificity.

H-FABP is a small protein belonging to the fatty acid-binding proteins (FABPs) and known to be expressed in the brain [31]. In myocardial infarction, HFABP is quickly released after the tissue damage [32,33]. It has been suggested that the release of H-FABP from damaged cells could be used for diagnosis of acute and chronic brain injuries [31]. GSTP-1 is a member of the Glutathione S-transferase superfamily, playing a role in oxidative stress. Its expression in brain has not been well studied, but GSTP-1 seems to be the main isoform in brain [34] and may function as a brain damage biomarker [24]. Our results showed a higher level of both H-FABP and GSTP-1 in CSF of late stage patients compared to early stage patients. These two molecules are known to be associated with early brain cell death [24,31], which could be correlated with the observed increase of their concentration in late-stage HAT patients. From now, it is not know if these two molecules were also associated with the inflammatory process.

Cytokines and chemokines play an important role in inflammatory processes and blood-brain barrier (BBB) dysfunction [35], and could therefore be potentially used as markers for staging HAT [11,16,36]. In the present study, the measured levels of inflammation-related proteins in CSF showed significant differences according to the disease progression. Indeed, concentrations of IL-1β, IL-6, IL-10, TNF-α, CXCL8 and CXCL10 were increased in the CSF of patients in late stage HAT compared to those in early stage of the disease. In addition, the levels of IL-1β, IL-6, CXCL8 and IL-10 were similar to those already reported for *T. b. gambiense* HAT [11,16]. IL-1β is a pro-inflammatory cytokine that induces leukocytes infiltration [37] and is rapidly expressed in response to brain damage [35]. The high level of IL-1β found in CSF of stage 2 patients confirmed its probable association with the inflammatory process. Furthermore, its level was clearly correlated to the presence of severe neurological signs, supporting a potential release in relation to neurodegeneration. IL-6 and IL-10 are both anti-inflammatory cytokines. Their increased level in the CSF according to the stage as well as the severity of the neurological signs confirmed their activation associated with disease progression. The concentration of the two molecules was significantly increased in patients with more than 20 WBC/μL, which may suggest a probable expression after an already activated inflammatory process. Indeed, it has been demonstrated in vervet monkey models of HAT that IL-10 is associated with down-regulation of pro-inflammatory cytokines (IFN-γ and TNF-α) in the late stage of *T. b. rhodesiense* disease [38]. The level of the pro-inflammatory chemokine CXCL8 was also significantly elevated in CSF of S2 patients and correlated well with both presence of trypanosomes in CSF and severity of neurological signs. CXCL8 is a strong neutrophil attractant [16], which could thus not explain the good correlation of CXCL8 and the number of WBC (mainly B-lymphocytes) in CSF. However, its elevation in patients with a relatively low number of WBC (between 5 and 20/μL) suggests an early activation, which may play a role in BBB dysregulation [11].

The pro-inflammatory cytokine TNF-α has been reported as being involved in blood-brain barrier dysfunction [39]. These authors also demonstrated that trypanosomes may induce synthesis of TNF-α. In the present study, the increasing level of TNF-α was associated with disease progression as well as the presence of the parasite in CSF. These results suggested that parasites invasion into the CNS may lead to TNF-α production, which generated then CNS inflammation [14]. Additionally, an elevation according to the severity of the neurological symptoms was observed, which may support the neurotoxic effect of this cytokine in HAT [35].

CXCL10, also known as IP-10, is a pro-inflammatory chemokine with a central role in inflammatory responses [40]. The main effect of CXCL10 as a chemotactic molecule is activation of T cell migration to the site of inflammation, after binding to its receptor, CXCR3 [41]. The involvement of this chemokine in different CNS disorders has been demonstrated, such as viral meningitis [42] and multiple sclerosis [43], where increased CXCL10 levels in the CSF correlated with tissue infiltration of T lymphocytes [44]. In our study, the concentration of CXCL10 increased with progression of the disease, and was highly correlated with the number of WBC in CSF. Many studies have pointed out astrocytes as the primary source of CXCL10 at the level of the CNS and showed that this molecule is responsible, as chemoattractant, for the influx of activated T lymphocytes in brain [43,45–47]. Indeed, there is a predominance of plasma cell infiltration in the brain of trypanosomiasis infected individuals. In addition, it has very recently been shown in a mouse model of HAT that CXCL10 may play an important role in T-cell recruitment into the brain parenchyma and is probably associated with brain invasion by trypanosomes [48]. Furthermore, the early activation of cytokine production (TNF-α, IL-6, and IFN-γ) by astrocytes and microglia in mice models infected with *T. brucei* before observation of an inflammatory response [49] has confirmed an important role of astrocyte activation in CNS inflammatory response. In consequence, early astrocyte activation, which induces CXCL10 production, is probably linked with BBB dysfunction and may occur before the inflammatory process. These hypotheses were supported by the increase CXCL10 concentration observed in patients having >5 and ≤ 20 WBC/μL but without trypanosomes detected in the CSF. The CXCL10 level was also demonstrated to be elevated in patients with cerebral malaria, and pointed out as potentially inducing apoptosis of endothelial cells leading to BBB breakdown [17] Recent work has suggested that neuronal apoptosis associated with calcium dysregulation may be induced by CXCL10 [50]. Even if mechanisms of CXCL10 mediated neurotoxicity remain unclear, we showed that the concentration of CXCL10 was correlated to the severity of neurological signs, supporting a possible involvement of this protein in neuronal injury pathways. Thus, CXCL10 expression in late stage HAT patients may be associated with both cell death and inflammatory process. Finally, active tuberculosis and pregnancy, two exclusion criteria in this study, have also been reported as modulating the level of CXCL10 [51,52]. Although they have only been evaluated on serum and whole blood samples so far, it is not excluded that these criteria could potentially induce CXCL10 modulation in CSF. Nevertheless, our data demonstrated that CXCL10 is an efficient tool for staging patients, and suggested a potential role of CXCL10 as an early marker of parasite invasion into the CNS.

As the investigated proteins may be involved in different biological mechanisms, we evaluated in this study a strategy to combine results of each molecule, in order to find a panel able to discriminate more accurately early and late stage patients. This highlighted a panel of three molecules, including CXCL10 (the most promising single molecule), CXCL8 (another chemokine) and H-FABP (a marker of brain damage). With a specificity of 100%, this panel increased the sensitivity for staging of HAT patients up to 97% (compared to the 84% obtained with CXCL10

taken individually). Although the number of "intermediate" patients was small, the panel appeared to classify them rather as S2 patients (7/8 patients). This supports the current recommendation by WHO to consider such patients as S2 patients and treat them with drugs used for late stage disease. However, there is a need for more studies on *T. b. gambiense* and *T. b. rhodesiense* patients, before and after treatment, as well as on other parasitic diseases such as cerebral malaria, to verify these results and assess the feasibility of using the three-molecule panel as a complement to WBC count. There are obviously some drawbacks concerning this approach. Firstly the obtained panel is not 100% sensitive and thus some stage 2 patients will not be detected. The influence of other possible co-infections should also be evaluated in order to determine if they significantly modulate the evaluated molecules. Indeed, the three molecules included in the panel could potentially all be markers of other CNS disorders. It is also evident that the methods described in this study could not be implemented in such a way directly in the field and should be first transformed into a more simplified technique as for example a lateral flow immunoassay. Another limitation is the continued requirement of the invasive lumbar puncture since the molecules highlighted in this study have been evaluated on CSF samples.

In conclusion, the present study demonstrated the utility of inflammation-related proteins and brain damage markers as indicators of the second stage of HAT but potentially in other CNS disorders as well. We highlighted the value of CXCL10 as an efficient staging biomarker for *T. b. gambiense* infected HAT patients. Additionally, a combination of CXCL10 with CXCL8 and H-FABP resulted in a highly sensitive tool for identification of late stage HAT patients.

Supporting Information

Figure S1 ROC curves of GR3 molecules and the panel. *Cut-off value for each molecule [pg/ml] and for the panel is displayed by a point and the numeric value. In parenthesis, sensitivity (%) of each molecule was set for 100% specificity. Area under the ROC curve (AUC) is also given.
Found at: doi:10.1371/journal.pntd.0000459.s001 (1.37 MB TIF)

Figure S2 Box-plot of GR3 molecules classified according to the stage of the disease. *Median and mean are represented as a solid line in the box and a diamond respectively. Whisks are defined as 5th–95th percentile without outliers. Half-width of the notch was calculated automatically by the software.
Found at: doi:10.1371/journal.pntd.0000459.s002 (0.69 MB TIF)

Table S1 Detailed results for GR3 molecules in function of the presence of trypanosomes in CSF (according or not to the stage) and the neurological signs.
Found at: doi:10.1371/journal.pntd.0000459.s003 (0.01 MB DOC)

Acknowledgments

The authors thank Karim Hammad for technical assistance and FIND for technical and scientific advice.

Author Contributions

Conceived and designed the experiments: VL JMN NT JCS. Performed the experiments: AH NT CF. Analyzed the data: AH NT XR FL MM NT JCS. Contributed reagents/materials/analysis tools: XR VL DMN EM JCE JCS. Wrote the paper: AH NT JCS.

References

1. World Health Organization (2006) Human African Trypanosomiasis (sleeping sickness): epidemiological update. Wkly Epidemiol Rec 81: 69–81.
2. Chappuis F, Loutan L, Simarro P, Lejon V, Büscher P (2005) Options for field diagnosis of human african trypanosomiasis. Clin Microbiol Rev 18: 133–146. doi:10.1128/CMR.18.1.133-146.2005.
3. Simarro PP, Jannin J, Cattand P (2008) Eliminating human African trypanosomiasis: where do we stand and what comes next? PLoS Med 5: e55. doi:10.1371/journal.pmed.0050055.
4. Lejon V, Büscher P (2005) Review Article: cerebrospinal fluid in human African trypanosomiasis: a key to diagnosis, therapeutic decision and post-treatment follow-up. Trop Med Int Health 10: 395–403. doi:10.1111/j.1365-3156.2005.01403.x.
5. Kennedy PGE (2006) Diagnostic and neuropathogenesis issues in human African trypanosomiasis. Int J Parasitol 36: 505–512. doi:10.1016/j.ijpara.2006.01.012.
6. Kennedy PGE (2008) The continuing problem of human African trypanosomiasis (sleeping sickness). Ann Neurol 64: 116–126. doi:10.1002/ana.21429.
7. Control and surveillance of African trypanosomiasis. Report of a WHO Expert Committee (1998) World Health Organ. Tech Rep Ser 881: I–VI, 1–114.
8. Lejon V, Buscher P (2001) Stage determination and follow-up in sleeping sickness. Med Trop (Mars) 61: 355–360.
9. Lejon V, Reiber H, Legros D, Djé N, Magnus E, et al. (2003) Intrathecal immune response pattern for improved diagnosis of central nervous system involvement in trypanosomiasis. J Infect Dis 187: 1475–1483. doi:10.1086/374645.
10. Kennedy PGE (2008) Diagnosing central nervous system trypanosomiasis: two stage or not to stage? Trans R Soc Trop Med Hyg 102: 306–307. doi:10.1016/j.trstmh.2007.11.011.
11. Lejon V, Lardon J, Kenis G, Pinoges L, Legros D, et al. (2002) Interleukin (IL)-6, IL-8 and IL-10 in serum and CSF of Trypanosoma brucei gambiense sleeping sickness patients before and after treatment. Trans R Soc Trop Med Hyg 96: 329–333. doi:10.1016/S0035-9203(02)90115-X.
12. Sternberg JM (2004) Human African trypanosomiasis: clinical presentation and immune response. Parasite Immunol 26: 469–476. doi:10.1111/j.0141-9838.2004.00731.x.
13. Masocha W, Rottenberg ME, Kristensson K (2007) Migration of African trypanosomes across the blood-brain barrier. Physiol Behav 92: 110–114. doi:10.1016/j.physbeh.2007.05.045.
14. Kennedy PGE (2009) Cytokines in central nervous system trypanosomiasis: cause, effect or both? Trans R Soc Trop Med Hyg 103: 213–214. doi:10.1016/j.trstmh.2008.08.013.
15. Maclean L, Odiit M, Macleod A, Morrison L, Sweeney L, et al. (2007) Spatially and genetically distinct African Trypanosome virulence variants defined by host interferon-gamma response. J Infect Dis 196: 1620–1628. doi:10.1086/522011.
16. Courtioux B, Boda C, Vatunga G, Pervieux L, Josenando T, et al. (2006) A link between chemokine levels and disease severity in human African trypanosomiasis. Int J Parasitol 36: 1057–1065. doi:10.1016/j.ijpara.2006.04.011.
17. Armah HB, Wilson NO, Sarfo BY, Powell MD, Bond VC, et al. (2007) Cerebrospinal fluid and serum biomarkers of cerebral malaria mortality in Ghanaian children. Malar J 6: 147. doi:10.1186/1475-2875-6-147.
18. John CC, Panoskaltsis-Mortari A, Opoka RO, Park GS, Orchard PJ, et al. (2008) Cerebrospinal fluid cytokine levels and cognitive impairment in cerebral malaria. Am J Trop Med Hyg 78: 198–205.
19. Jain V, Armah HB, Tongren JE, Ned RM, Wilson NO, et al. (2008) Plasma IP-10, apoptotic and angiogenic factors associated with fatal cerebral malaria in India. Malar J 7: 83. doi:10.1186/1475-2875-7-83.
20. Hu S, Loo JA, Wong DT (2006) Human body fluid proteome analysis. Proteomics 6: 6326–6353. doi:10.1002/pmic.200600284.
21. Lescuyer P, Allard L, Zimmermann-Ivol CG, Burgess JA, Hughes-Frutiger S, et al. (2004) Identification of post-mortem cerebrospinal fluid proteins as potential biomarkers of ischemia and neurodegeneration. Proteomics 4: 2234–2241. doi:10.1002/pmic.200300822.
22. Zimmermann-Ivol CG, Burkhard PR, Le Floch-Rohr J, Allard L, Hochstrasser DF, et al. (2004) Fatty acid binding protein as a serum marker for the early diagnosis of stroke: a pilot study. Mol Cell Proteomics 3: 66–72. doi:10.1074/mcp.M300066-MCP200.
23. Guillaume E, Zimmermann C, Burkhard PR, Hochstrasser DF, Sanchez J (2003) A potential cerebrospinal fluid and plasmatic marker for the diagnosis of Creutzfeldt-Jakob disease. Proteomics 3: 1495–1499. doi:10.1002/pmic.200300478.
24. Burgess JA, Lescuyer P, Hainard A, Burkhard PR, Turck N, et al. (2006) Identification of brain cell death associated proteins in human post-mortem cerebrospinal fluid. J Proteome Res 5: 1674–1681. doi:10.1021/pr060160v.
25. Marchi N, Cavaglia M, Fazio V, Bhudia S, Hallene K, et al. (2004) Peripheral markers of blood-brain barrier damage. Clin Chim Acta 342: 1–12. doi:10.1016/j.cccn.2003.12.008.
26. Bloomfield S, McKinney J, Smith L, Brisman J (2007) Reliability of S100B in predicting severity of central nervous system injury. Neurocrit Care 6: 121–138. doi:10.1007/s12028-007-0008-x.
27. Miézan TW, Meda HA, Doua F, Djè NN, Lejon V, et al. (2000) Single centrifugation of cerebrospinal fluid in a sealed pasteur pipette for simple, rapid and sensitive detection of trypanosomes. Trans R Soc Trop Med Hyg 94: 293.

28. Allard L, Turck N, Burkhard PR, Walter N, Rosell A, et al. (2007) UFD1 as blood marker for the early diagnosis of stroke. Biomark Insights 2: 155–164.
29. R Development Core Team (2008) R: A Language and Environment for Statistical Computing. Vienna, Austria. Available: http://www.R-project.org.
30. Reynolds MA, Kirchick HJ, Dahlen JR, Anderberg JM, McPherson PH, et al. (2003) Early biomarkers of stroke. Clin Chem 49: 1733–1739. doi:10.1373/49.10.1733.
31. Lescuyer P, Allard L, Hochstrasser DF, Sanchez J (2005) Heart-fatty acid-binding protein as a marker for early detection of acute myocardial infarction and stroke. Mol Diagn 9: 1–7.
32. Glatz JF, van Bilsen M, Paulussen RJ, Veerkamp JH, van der Vusse GJ, et al. (1988) Release of fatty acid-binding protein from isolated rat heart subjected to ischemia and reperfusion or to the calcium paradox. Biochim Biophys Acta 961: 148–152. doi:10.1016/0005-2760(88)90141-5.
33. Knowlton AA, Apstein CS, Saouf R, Brecher P (1989) Leakage of heart fatty acid binding protein with ischemia and reperfusion in the rat. J Mol Cell Cardiol 21: 577–583. doi:10.1016/0022-2828(89)90823-7.
34. Theodore C, Singh SV, Hong TD, Awasthi YC (1985) Glutathione S-transferases of human brain. Evidence for two immunologically distinct types of 26500-Mr subunits. Biochem J 225: 375–382.
35. Allan SM, Rothwell NJ (2001) Cytokines and acute neurodegeneration. Nat Rev Neurosci 2: 734–744. doi:10.1038/35094583.
36. Sternberg JM, Rodgers J, Bradley B, Maclean L, Murray M, et al. (2005) Meningoencephalitic African trypanosomiasis: brain IL-10 and IL-6 are associated with protection from neuro-inflammatory pathology. J Neuroimmunol 167: 81–89. doi:10.1016/j.jneuroim.2005.06.017.
37. Ching S, He L, Lai W, Quan N (2005) IL-1 type I receptor plays a key role in mediating the recruitment of leukocytes into the central nervous system. Brain Behav Immun 19: 127–137. doi:10.1016/j.bbi.2004.06.001.
38. Ngotho M, Maina N, Kagira J, Royo F, Farah IO, et al. (2006) IL-10 is up regulated in early and transitional stages in vervet monkeys experimentally infected with Trypanosoma brucei rhodesiense. Parasitol Int 55: 243–248. doi:10.1016/j.parint.2006.06.004.
39. Girard M, Giraud S, Courtioux B, Jauberteau-Marchan M, Bouteille B (2005) Endothelial cell activation in the presence of African trypanosomes. Mol Biochem Parasitol 139: 41–49. doi:10.1016/j.molbiopara.2004.09.008.
40. Huang D, Han Y, Rani MR, Glabinski A, Trebst C, et al. (2000) Chemokines and chemokine receptors in inflammation of the nervous system: manifold roles and exquisite regulation. Immunol Rev 177: 52–67. doi:10.1034/j.1600-065X.2000.17709.x.
41. Weng Y, Siciliano SJ, Waldburger KE, Sirotina-Meisher A, Staruch MJ, et al. (1998) Binding and functional properties of recombinant and endogenous CXCR3 chemokine receptors. J Biol Chem 273: 18288–18291.
42. Lahrtz F, Piali L, Nadal D, Pfister HW, Spanaus KS, et al. (1997) Chemotactic activity on mononuclear cells in the cerebrospinal fluid of patients with viral meningitis is mediated by interferon-gamma inducible protein-10 and monocyte chemotactic protein-1. Eur J Immunol 27: 2484–2489. doi:10.1002/eji.1830271004.
43. Sørensen TL, Tani M, Jensen J, Pierce V, Lucchinetti C, et al. (1999) Expression of specific chemokines and chemokine receptors in the central nervous system of multiple sclerosis patients. J Clin Invest 103: 807–815. doi:10.1172/JCI5150.
44. Dufour JH, Dziejman M, Liu MT, Leung JH, Lane TE, et al. (2002) IFN-gamma-inducible protein 10 (IP-10; CXCL10)-deficient mice reveal a role for IP-10 in effector T cell generation and trafficking. J Immunol 168: 3195–3204.
45. Farina C, Krumbholz M, Giese T, Hartmann G, Aloisi F, et al. (2005) Preferential expression and function of Toll-like receptor 3 in human astrocytes. J Neuroimmunol 159: 12–19. doi:10.1016/j.jneuroim.2004.09.009.
46. Hanum PS, Hayano M, Kojima S (2003) Cytokine and chemokine responses in a cerebral malaria-susceptible or -resistant strain of mice to Plasmodium berghei ANKA infection: early chemokine expression in the brain. Int Immunol 15: 633–640.
47. van Heteren JT, Rozenberg F, Aronica E, Troost D, Lebon P, et al. (2008) Astrocytes produce interferon-alpha and CXCL10, but not IL-6 or CXCL8, in Aicardi-Goutières syndrome. Glia 56: 568–578. doi:10.1002/glia.20639.
48. Amin DN, Rottenberg ME, Thomson AR, Mumba D, Fengers C, et al. (2009) Expression and role of CXCL10 during the encephalitic stage of experimental and clinical African trypanosomiasis. Proc Natl Acad Sci U S A;In press.
49. Hunter CA, Jennings FW, Kennedy PG, Murray M (1992) Astrocyte activation correlates with cytokine production in central nervous system of Trypanosoma brucei brucei-infected mice. Lab Invest 67: 635–642.
50. Sui Y, Stehno-Bittel L, Li S, Loganathan R, Dhillon NK, et al. (2006) CXCL10-induced cell death in neurons: role of calcium dysregulation. Eur J Neurosci 23: 957–964. doi:10.1111/j.1460-9568.2006.04631.x.
51. Whittaker E, Gordon A, Kampmann B (2008) Is IP-10 a better biomarker for active and latent tuberculosis in children than IFNgamma? PLoS ONE 3: e3901. doi:10.1371/journal.pone.0003901.
52. Gotsch F, Romero R, Friel L, Kusanovic JP, Espinoza J, et al. (2007) CXCL10/IP-10: a missing link between inflammation and anti-angiogenesis in preeclampsia? J Matern Fetal Neonatal Med 20: 777–792. doi:10.1080/14767050701483298.

7

Discussion, conclusions and perspectives

In this thesis, we investigated the feasibility of the combination of biomarkers into panels. We especially focused our work on the definition of clear and understandable models, and their validation with statistical means.

After a general introduction of the methods available for combining biomarkers (chapter 2), we presented the two tools developed during this project. The pROC package for R and S+ (chapter 3) is dedicated to ROC analysis. It features several ROC comparison tests and other statistical methods that are not available together in most statistical software. PanelomiX (chapter 4) is a workflow to combine biomarkers based on thresholds with a web interface. Finally, chapters 5 and 6 presented two clinical applications with aneurysmal subarachnoid hemorrhage and human African trypanosomiasis.

This concluding chapter summarizes and discusses the results presented in the papers in the context of the goals of this thesis, and proposes a few possibilities to build better biomarker panels.

1 *Propose a framework to easily create white-box panels of biomarkers*

The first goal of this thesis was to explore ways to combine biomarkers into panels. We investigated several established methods, such as logistic regression, decision trees and support vector machines.

We also implemented an approach based on thresholds where the score is computed as the sum of positive biomarkers. We created a tool, PanelomiX, that determines the set of biomarkers to be included in the panel and computes the thresholds associated with the markers at the same time, corresponding to an embedded multivariate feature selection. To generate an optimal classification, it is performed through an exhaustive search.

Due to the computational complexity of the exhaustive search, this approach is not applicable for datasets where more than about 10 biomarkers must be combined. Therefore, we investigated several alternative methods and feature selection methods. We found that Random Forest, a combination method based on decision trees and bootstrapping, could be efficiently employed to outline the most

interesting biomarkers and thresholds. This pre-filtering method shows very interesting multivariate characteristics, but being based on decision tree, the optimal feature set can be very different from the set that would be optimal for threshold-based combinations. In addition, correlated features may be rated with lower frequency[1] and could thus be spuriously rejected, whereas one of the features in the correlated set could have been selected with the fully exhaustive search.

Other methods were tested such as the top-down and bottom-up approaches proposed by Calzolari *et al.*[2] and genetic algorithms[3]. Like the exhaustive search, these methods can be employed to determine simulataneously both the set of markers and the associated thresholds, thus representing powerful multivariate methods with embedded feature selection.

The method proposed by Calzolari *et al.* was originally designed to test drug combinations in *in vivo* studies where each iteration can take up to several weeks[2] . It is very fast, but it can easily be trapped into local maxima. Indeed, tests showed that panels fitted with this method displayed much less accuracy than with exhaustive search and random forest pre-processing. Genetic algorithms were also tested. The main issue we faced was that both the set of biomarkers and the thresholds must be optimized. Most current implementations of genetic algorithms optimize one or more numeric values but are not able to select which variables must be included. This is for instance the case of the *genalg* R package[4]. While the *subselect* package[5] can theoretically do it, it works on correlation matrices rather than on the data itself. It uses generalized or multivariate linear models and therefore does not take into account the interactions that are specific to the threshold classification method.

Further improvements are still possible. First, the exhaustive search was parallelized. However, tests with highly parallel machines showed that it currently does not run on more than about 4 cores. Second, the selection of the optimal panel could be improved with approaches such as Akaike information criterion (AIC) or Bayesian information criterion (BIC)[6] that takes into account the number of biomarkers included in the panel to favor panels with less biomarkers. It is typically applied with log-likelihood estimates of the performance, which could be developed in the context of threshold combinations.

2 *Study the performance of the proposed panels, in comparison with single biomarkers and other established methods.*

The second objective of the project was to estimate how this white-box panel method compared with individual biomarkers in order to get a better estimation of its true benefits. We implemented an approach based on receiver operating characteristic (ROC) analysis, with tests to compare full or partial areas under the ROC curve. As the performance of the panels is overestimated when it is measured on the same data that was employed to train it, and because independent test sets are most often not available in clinical research, we employed cross-validation to obtain an estimate of the performance of the panels free of over-fitting. In order to perform a fair comparison, and because single biomarkers have been shown to be over-fitted too[7], we had to find a way to cross-validate the biomarkers. Because single markers are often analyzed with a threshold, we applied it in a cross-validation setup. The result is a panel containing one biomarker, and a ROC curve with one single point in the ROC space. It has been shown before that the area under this kind of ROC curve is negatively biased[8]. However, this effect can be mitigated. To this end, we repeated the cross-validation several times, and averaged the predictions obtained over the runs. The resulting ROC curve is smoothed around the point of sensitivity/specificity of interest, and the comparison of the partial AUC is then valid.

The comparison with other combination methods is straightforward, because the cross-validation can be applied in exactly the same way as for the PanelomiX panel. Therefore, the ROC estimates can be directly compared.

As expected, we found that the performance of PanelomiX compared favorably with the separate biomarkers, even without the cross-validation (chapter 4). In addition, PanelomiX also yielded a better classification than established methods like SVM, Rpart and logistic regression on the aneurysmal subarachnoid haemorrhage dataset.

Nevertheless, this approach suffers from several limitations. First, the ROC analysis and especially the AUC do not give precise information about the sensitivity and specificity of the test and therefore does not precisely indicate how well the patients of each class will be classified. An alternative to ROC analysis is to compute

contingency tables and apply McNemar's test [9]. However, this test suffers from several limitations too[10], the most critical of which is its sensitivity to the proportion of negative to positive cases. This could be addressed by considering only diseased patients for the comparison of sensitivities, and healthy patients for the specificities[11]. Second, the validity of the method of comparison between ROC curves of panels with single markers should be assessed through simulation to ensure that the comparison is really unbiased.

3 *Build interfaces to be used by the scientists in the lab*

As it was performed embedded in a mostly "wet"-lab, one of the main targets of the project was to produce tools that can be easily used by researchers without an extended knowledge in bioinformatics. They need to be able to operate the tools by themselves, without the help of a bioinformatician. Therefore, a programming library or a command-line tool is not acceptable and a graphical user interface (GUI), potentially web-based, is required.

The first tool that was built and publicly released is pROC. As presented in chapter 3, it includes both a GUI for non-programmer users, and a command-line interface in R and S+ for the users of those languages and statistical environments. As it was released on the CRAN, the public repository for R packages, it quickly gained in popularity and was already used in more than 30 published research articles in various international groups. To take only a few examples, McLaughlin employed it to propose stressor-response model in water quality management[12], and Leichtle *et al.* applied it to the evaluation of panels of metabolites[13]. Bryceson *et al.* compared the performance of several assays[14], Einav *et al.* evaluated the performance of two biomarkers[15], Ignatiadis *et al.* predicted the outcome after breast cancer[16] and Plaisier *et al.* selected models of miRNA regulatory networks[17].

The second tool was PanelomiX. It is described in detail in chapter 4, together with the algorithms implemented. It will also be published soon. It features a web-based interface for an easy management of panels, from data submission to results display. Even though it has not been the subject of a publication yet, it has been applied to

various clinical datasets by three scientists of the lab and will be commercialized by a British private company.

4 Applicability to other kinds of datasets

All the applications shown in chapters 3 to 6 were datasets available in the lab. Proteins were already measured with ELISA or equivalent techniques, and the combination of biomarkers was an additional step in the analysis that wasn't planned at the time of the collection of the data. We also applied this methodology to two third-party datasets. The results are briefly described and commented in the next few pages.

4.1 Alzheimer's disease study

A partner from Kings College in London developed SRM assays for several proteins of interest in the diagnosis of Alzheimer's disease. Nine proteins were represented by one or more peptides, with one or more transitions measured for each peptide on two different instruments, a triple quadrupole and an ion trap. Quantitative data was available for 89 patients, 29 controls and 60 patients suffering from Alzheimer's disease. We first carried out a univariate analysis with pROC to highlight the most interesting transitions. Striking differences were observed between the two instruments. For instance on the triple quadrupole, a transition of the Complement C3 protein was found with 97% specificity (95% CI: 90-100%) and 29% sensitivity (95% CI: 17-41%). On the ion trap, the same transition displayed only 10% sensitivity (95% CI: 3-18%) at a similar specificity. On the other hand, the transition with the best discrimination power on the ion trap was one of the Complement factor H that showed a sensitivity of 18% (95% CI: 10-28%) with 97% specificity (95% CI: 90-100%), but this transition had less than 7% sensitivity (95% CI: 0-42%) on the triple quadrupole. Confidence intervals were rather large due to the rather limited number of patients and the low signal of the transitions, and the differences are not significant. Nonetheless, this analysis highlighted the potential utility of PanelomiX as an additional first filter to select SRM transitions[18].

Next, we applied PanelomiX to find the best combinations of 3 biomarkers. The data was analyzed in two ways: transitions were first analyzed as separate predictors (40

variables), and then we took the median of all the transitions of a protein as predictor (resulting in 9 variables). The rationale of the averaging is to reduce the noise present in the data. This would work well if the transitions have similar ionization and flying properties in the mass spectrometer. However, it is likely not the case due to post-translational modifications that affect the peptides specifically, and different peptides could have different performance characteristics. Some transitions of a given peptide may also display higher classification performances, especially if they have better ionization characteristics. In this case, averaging the transitions would reduce the classification performance of the data. Indeed, we observed the best classification performance with the separate transitions.

On the TSQ dataset, we found a panel of 3 transitions with 97% specificity (95% CI: 90-100%) and 54% (95% CI: 42-67%) sensitivity on the training set. However, this panel was not confirmed with cross-validation where only one out of the 3 transitions was consistently selected through the folds. Many transitions could replace the 2 other ones as the dataset slightly changes. In addition, with 40 variables and 89 patients, the level of over-fitting was large and the performance on the test sets was poor. Random Forest pre-processing was also applied. The level of over-fitting was slightly reduced, but the fitting itself was reduced too (46% specificity, 95% CI: 33-60%, instead of 54%) and the performance on cross-validation was poor.

Overall, this approach could highlight a very promising transition of the Complement C3 protein, that appeared superior to the other ones both with separate ROC analysis and in combination. However, the sample size was too small to draw robust conclusions. In addition, as shown in chapter 4, panels of 3 biomarkers does not improve the classification as much as panels combining more markers. A larger dataset with 900 patients is currently being analyzed with PanelomiX and should bring much more robust results.

4.2 The breast cancer study

The second dataset on which we applied PanelomiX was a microarray study on breast cancer published in 2002[19,20]. The goal was to predict the development of

distant metastases within 5 years. To that end, the expression of 19553 genes was measured for 78 training patients.

Because of the high dimensionality of this dataset, Random Forest pre-processing was applied. When all the genes were analyzed, we figured out that Random Forest was not able to cope with so many variables. Indeed, computation time increases exponentially with the number of variables. It took about one minute to compute the forest with 1000 variables on a desktop computer, and doing it for all genes was not realistic. Therefore, we limited the Random Forest to the 187 genes correlated with the outcome as was done by the authors. It should be noted that this univariate filter may well reject proteins that would be interesting in a panel. More advanced methods such as clustering could probably generate better results.

Next, we analyzed the stability of the Random Forest pre-filtering by repeating 1000 times the selection of 10 genes. No gene was selected in more than 12% of the runs, and no gene was never selected. This indicates that most genes can be substituted each other and their presence is not critical in panels. This finding is not surprising as all those genes were selected because they were correlated with the outcome of interest.

Finally, we ran PanelomiX on this dataset, searching for panels of at most 3 genes. We found panels with good performances, but they were not confirmed when applied to the validation set of 295 partially-redundant patients where the performance was not higher than that of single genes.

In conclusion, PanelomiX was able to deal with different types of datasets. These studies highlighted several limitations of PanelomiX, opening new perspectives for the improvement of the method. For example, in the case where more than about one thousand biomarkers are available, they must be filtered externally because Random Forest is not able to deal with it.

5 *Perspectives*

In the future, biomarkers research will need to think about combinations already during the discovery phase. A recent paper by Brasier *et al.* carried out such a combination-aware biomarker discovery[21]. To diagnose dengue hemorrhagic fever

they combined 34 proteins found significantly differentially expressed with 2-DE and LC-MS/MS and 2 cytokines. They directly applied a multivariate adaptive regression splines algorithm and found a panel of 8 biomarkers: one cytokine and 7 2-DE spots. This study shows an interesting approach to the combination of biomarkers. However, they applied a first univariate filter with statistical significance that could have rejected interesting biomarkers with only a small discrimination power that failed the significance test but could prove more significant when combined with other biomarkers.

In the same manner, all the biomarkers that have been tested during this project have passed a first univariate filter. Indeed, the biomarker discovery phase focused only on the individual performance of the biomarkers, and combinations were never considered in that phase. In the trypanosomiasis staging project presented in chapter 6, the biomarkers were chosen based on literature research and the *post-mortem* cerebrospinal fluid approach[22,23]. For the determination of the outcome 6 month after aneurysmal subarachnoid hemorrhage, presented in chapter 5, the biomarkers were either based also on the *post-mortem* cerebrospinal fluid model (H-FABP, NDKA, UFD-1) or biomarkers that were already measured in the hospital (S100b, WFNS, Fisher, Troponin I). Again, this univariate filter may have rejected biomarkers that did not have a high performance when they were taken individually, but that could have a significant impact in a panel.

Therefore, it is crucial to develop multivariate approaches to the early discovery of biomarkers. Unfortunately, PanelomiX is not yet appropriate to compute panels on datasets with many proteins such as protein microarray datasets. Random Forest was able to deal with up to 1000 biomarkers approximately, which is sufficient for most current discovery projects based on mass spectrometry, but not with the thousands of proteins that can be measured in protein arrays that must be pre-filtered. The efficiency of methods such as shrunken centroids[24], correlation-based feature selection[25], Markov blanket filters[26] or other common feature selection methods for microarray datasets[27] should be investigated in this context. In addition the reliability of this approach must still be assessed in such setups with simulation studies.

Another promising research area is the network biomarkers approach. It has been shown[28] that it was essential to study perturbations and responses in the context of the whole cell's network. For instance, Janes *et al.* showed that the JNK kinase can be seen both as pro- or anti-apoptotic, depending on the state of the phosphorylation network[29]. While this study focused on kinases in a cell context, it can hold true for biomarkers where the level of a biomarker is influenced by the multivariate state of the organism. With such an approach, a biomarker can be not only the state of the body at a given time, but also correlations (or lack of correlations) in the system[30]. Network biomarker approaches, by improving our knowledge of cell function, could highlight new potential biomarker targets.

6 Conclusion

Panels of biomarkers represent a promising way for potential improvements in the classification of patients in clinical studies. They combine information from several biomarkers into a single output with improved performance characteristics. Nevertheless, they come with specific drawbacks. First, the higher risk of overfitting compared with simgle biomarkers requires a higher number of patients and specific validation schemes to verify the performance claims. Second, the higher costs caused by the higher number of measurements requires that the cost to efficiency ratio be carefully studied.

In this thesis we focused on the former issue. We implemented an approach to combine biomarkers based on threshold decisions, called PanelomiX. We showed that it was possible to study the performance of such classifiers with receiver-operating characteristic (ROC) analysis and areas under the curve (AUC). We implemented pROC, a tool to carry this kind of analysis. We finally applied it to compare the panel with separate biomarkers and classical classification methods.

This work represents a preliminary attempt to approach the subject of biomarker combinations. While similar approaches have been described in the literature, they often lack important validation procedures. Future developments include embedding the multivariate dimension during the biomarker discovery phase and

the discovery of new potential panels of biomarkers with network biology approaches.

7 References

1. Nicodemus K. K., Malley J. D., Strobl C., *et al.*, (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11 (1), p. 110. DOI: 10.1186/1471-2105-11-110.

2. Calzolari D., Bruschi S., Coquin L., *et al.*, (2008). Search Algorithms as a Framework for the Optimization of Drug Combinations. *PLoS Computational Biology*, 4 (12), p. e1000249. DOI: 10.1371%2Fjournal.pcbi.1000249.

3. Goldberg D. E., (1989). *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley.

4. Willighagen E., (2005). *genalg: R Based Genetic Algorithm*,

5. Cerdeira J. O., Silva P. D., Cadima J., *et al.*, (2011). *subselect: Selecting variable subsets*,

6. Hastie T., Tibshirani R. & Friedman J., (2003). *Elements of Statistical Learning: data mining, inference, and prediction* Springer-Verlag., New York.

7. Whiteley W., Chong W. L., Sengupta A., *et al.*, (2009). Blood Markers for the Prognosis of Ischemic Stroke: A Systematic Review. *Stroke*, 40 (5), p. e380-389. DOI: 10.1161/STROKEAHA.108.528752.

8. Hanley J. A. & McNeil B. J., (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143 (1), p. 29-36.

9. McNemar Q., (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12 (2), p. 153-157. DOI: 10.1007/BF02295996.

10. Hawass N. E., (1997). Comparing the Sensitivities and Specificities of Two Diagnostic Procedures Performed on the Same Group of Patients. *British Journal of Radiology*, 70 (832), p. 360-366.

11. Trajman A. & Luiz R. R., (2008). McNemar χ^2 test revisited: comparing sensitivity and specificity of diagnostic examinations. *Scandinavian Journal of Clinical & Laboratory Investigation*, 68 (1), p. 77-80. DOI: 10.1080/00365510701666031.

12. McLaughlin D. B., Assessing the predictive performance of risk-based water quality criteria using decision error estimates from ROC analysis. *Integrated Environmental Assessment and Management*, accepted. DOI: 10.1002/ieam.1301.

13. Leichtle A. B., Nuoffer J.-M., Ceglarek U., *et al.*, (2012). Serum amino acid profiles and their alterations in colorectal cancer. *Metabolomics*, 8 (4), p. 643-653. DOI: 10.1007/s11306-011-0357-5.

14. Bryceson Y. T., Pende D., Maul-Pavicic A., *et al.*, (2012). A Prospective Evaluation of Degranulation Assays in the Rapid Diagnosis of Familial Hemophagocytic Syndromes. *Blood*, 119 (12), p. 2754-2763. DOI: 10.1182/blood-2011-08-374199.

15. Einav S., Kaufman N., Algur N., *et al.*, (2012). Modeling Serum Biomarkers S100 Beta and Neuron-Specific Enolase as Predictors of Outcome After Out-of-Hospital Cardiac Arrest. *Journal of the American College of Cardiology*, 60 (4), p. 304-311. DOI: 10.1016/j.jacc.2012.04.020.

16. Ignatiadis M., Singhal S. K., Desmedt C., *et al.*, (2012). Gene Modules and Response to Neoadjuvant Chemotherapy in Breast Cancer Subtypes: A Pooled Analysis. *Journal of Clinical Oncology*, 30 (16), p. 1996-2004. DOI: 10.1200/JCO.2011.39.5624.

17. Plaisier C. L., Pan M. & Baliga N. S., (2012). A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. *Genome Research*. DOI: 10.1101/gr.133991.111.

18. Lange V., Picotti P., Domon B., *et al.*, (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular Systems Biology*, 4 (1). DOI: 10.1038/msb.2008.61.

19. van de Vijver M. J., He Y. D., van't Veer L. J., *et al.*, (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347 (25), p. 1999-2009. DOI: 10.1056/NEJMoa021967.

20. van 't Veer L. J., Dai H., van de Vijver M. J., *et al.*, (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415 (6871), p. 530-536. DOI: 10.1038/415530a.

21. Brasier A. R., Garcia J., Wiktorowicz J. E., *et al.*, (2012). Discovery Proteomics and Nonparametric Modeling Pipeline in the Development of a Candidate Biomarker Panel for Dengue Hemorrhagic Fever. *Clinical and Translational Science*, 5 (1), p. 8-20. DOI: 10.1111/j.1752-8062.2011.00377.x.

22. Lescuyer P., Allard L., Zimmermann-Ivol C. G., *et al.*, (2004). Identification of post-mortem cerebrospinal fluid proteins as potential biomarkers of ischemia and neurodegeneration. *PROTEOMICS*, 4 (8), p. 2234-2241. DOI: 10.1002/pmic.200300822.

23. Burgess J. A., Lescuyer P., Hainard A., *et al.*, (2006). Identification of Brain Cell Death Associated Proteins in Human Post-mortem Cerebrospinal Fluid. *Journal of Proteome Research*, 5 (7), p. 1674-1681. DOI: 10.1021/pr060160v.

24. Tibshirani R., Hastie T., Narasimhan B., *et al.*, (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99 (10), p. 6567-6572. DOI: 10.1073/pnas.082099299.

25. Wang Y., Tetko I. V., Hall M. A., *et al.*, (2005). Gene selection from microarray data for cancer classification—a machine learning approach. *Computational Biology and Chemistry*, 29 (1), p. 37-46. DOI: 10.1016/j.compbiolchem.2004.11.001.

26. Zhu Z., Ong Y.-S. & Dash M., (2007). Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, 40 (11), p. 3236-3248. DOI: 10.1016/j.patcog.2007.02.007.

27. Lazar C., Taminau J., Meganck S., *et al.*, (2012). A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9 (4), p. 1106-1119. DOI: 10.1109/TCBB.2012.33.

28. Erler J. T. & Linding R., (2010). Network-based drugs and biomarkers. *The Journal of Pathology*, 220 (2), p. 290-296. DOI: 10.1002/path.2646.

29. Janes K. A., Albeck J. G., Gaudet S., *et al.*, (2005). A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis. *Science*, 310 (5754), p. 1646-1653. DOI: 10.1126/science.1116598.

30. Taylor I. W., Linding R., Warde-Farley D., *et al.*, (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, 27 (2), p. 199. DOI: 10.1038/nbt.1522.

List of publications

This thesis resulted in the publication of the following articles in peer-reviewd journals:

1. Robin X., Turck N., Hainard A., *et al.*, (2009). Bioinformatics for protein biomarker panel classification: what is needed to bring biomarker panels into in vitro diagnostics? *Expert Review of Proteomics*, 6 (6), p. 675-689. DOI: 10.1586/epr.09.83.

2. Hainard A., Tiberti N., Robin X., *et al.*, (2009). A Combined CXCL10, CXCL8 and H-FABP Panel for the Staging of Human African Trypanosomiasis Patients. *PLoS Neglected Tropical Diseases*, 3 (6), p. e459. DOI: 10.1371/journal.pntd.0000459.

3. Tiberti N., Hainard A., Lejon V., *et al.*, (2010). Discovery and Verification of Osteopontin and Beta-2-microglobulin as Promising Markers for Staging Human African Trypanosomiasis. *Molecular & Cellular Proteomics*, 9 (12), p. 2783-2795. DOI: 10.1074/mcp.M110.001008.

4. Turck N., Vutskits L., Sanchez-Pena P., *et al.*, (2010). A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Medicine*, 36 (1), p. 107-115. DOI: 10.1007/s00134-009-1641-y.

5. Robin X., Turck N., Hainard A., *et al.*, (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, p. 77. DOI: 10.1186/1471-2105-12-77.

6. Hainard A., Tiberti N., Robin X., *et al.*, (2011). Matrix metalloproteinase-9 and intercellular adhesion molecule 1 are powerful staging markers for human African trypanosomiasis. *Tropical Medicine and International Health*, 16 (1), p. 119-126. DOI: 10.1111/j.1365-3156.2010.02642.x.

7. Tiberti N., Hainard A., Lejon V., *et al.*, (2012). Cerebrospinal Fluid Neopterin as Marker of the Meningo-Encephalitic Stage of Trypanosoma brucei gambiense Sleeping Sickness. *PLoS ONE*, 7 (7), p. e40909. DOI: 10.1371/journal.pone.0040909.

8. Turck N., Robin X., Walter N., et al., (2012). Blood Glutathione S-Transferase-π as a time predictor of stroke onset. PLoS ONE, accepted.

In addition, the following manuscripts are in preparation:

1. Robin X., Turck N., Hainard A., *et al.* PanelomiX: a web-based tool to create biomarker panels based on thresholds. Manuscript in preparation.

2. Tiberti N., Lejon V., Hainard A., *et al.* Neopterin Is a New Cerebrospinal Fluid Marker for Treatment Outcome Evaluation in Patients Affected by Trypanosoma brucei gambiense Sleeping Sickness . Manuscript in preparation.

3. Tiberti N., Matovu E., Hainard A., *et al.* New biomarkers for stage determination in patients affected by Trypanosoma brucei rhodesiense sleeping sickness . Manuscript in preparation.

4. Walder B., Robin X., My Lien Rebetez M., *et al.* The diagnostic and prognostic significance of the serum biomarkers H-FABP in comparison with S-100b in severe traumatic brain injury. Manuscript in preparation.

5. Copin J.-C., My Lien Rebetez M., Turck N., *et al.* MMP-9 and cellular fibronectin plasma concentrations are predictor of the hospital length of stay after severe traumatic brain injury. Manuscript in preparation.