A multivariate analysis of protein microarrays for signature selection profiles

Saveria Mazzara#✉, Antonella Sinisi, Angela Cardaci, Sergio Abrignani, Mauro Bombaci#✉

Istituto Nazionale Genetica Molecolare (INGM), Milan, Italy

# These authors have contributed equally to this work

Motivations and Objectives

In recent years, protein microarrays have become one of the most invaluable research tools in the field of large-scale and high-throughput biology, and their use in basic research, diagnostics and drug discovery has emerged as a great promise of medicine. An interesting application of this technology is the identification of a serodiagnostic antigens ensemble whose expression profiles can effectively unveil discriminant patterns providing the classification of healthy and disease samples.

Nowadays, the analysis of protein microarray data for extracting biologically interpretable results is still an extremely complex process, and there is an increasing need for fully automated data mining approaches.

In the present study a Partial Least Squares Discriminant Analysis (PLS-DA) has been applied to protein microarrays aimed to reveal discriminative patterns between different clinical conditions. The method was evaluated to data generated from protein microarrays, including 1626 human recombinant proteins, probed with sera of patients with autoimmune liver diseases. Each array was depicted as a set of quantitative descriptors and analyzed by PLS-DA method in an attempt to classify samples according to their intrinsic protein expression profile. Moreover, the assessed model was able to extract antigens of interest representative of a different protein profile in Autoimmune Hepatitis (AIH) patients compared to Healthy donors (HD) (Zingaretti et al., 2012).

Here, the application of multivariate statistical techniques to protein microarray data represents an effective tool to identify informative protein profiles as of fully automatic strategy.

This kind of approach could lead to a more rapid and accurate development of diagnostic tests, providing useful factors able to discriminate different autoimmune diseases.

Methods

We proposed an innovative bioinformatic workflow, based on multivariate data analysis, for identifying discriminative patterns between different clinical conditions. A schematic view of flow chart is shown in Figure 1. In the first step (Figure 1, panel A), protein arrays were developed to screen serum samples of patients affected by Autoimmune Hepatitis (AIH) and healthy controls (HD); characteristic of patients and protein platform generation were described in recent publication (Zingaretti et al., 2012). Briefly, fluorescence signals were detected by using a ScanArray Gx PLUS (PerkinElmer, Bridgeport Avenue Shelton, USA) and scanned images were imported in a house developed software for the successive image analysis. Normalization to the spotted human IgG curve was performed (Bombaci et al., 2009). Subsequently, the data were analyzed by partial least squares discriminant analysis (PLS-DA) (Wold et al., 2001; Eriksson et al., 2006) (Figure 1, panel B) with the aim of identifying the best candidates for the development of new application in clinical research (Figure 1, panel C). In recent years, projection methods are being successfully applied to biological data such as DNA microarrays and proteomic data but the combination of PLS-DA with the protein arrays represents a new and interesting approach for investigation of this type of proteomics data. The method is particularly suitable for analysis of data with numerous variables and is able to integrate information about the response matrix, Y, into the descriptor matrix, X (the antigens). PLS method is based on finding the latent variables that maximize the covariance between X and Y. The importance of each variable in the loadings of PLS-DA is given by the variable influence on projection (VIP) parameter. The VIP score reflects the influence of antigens on the classification, and predictors with score larger than one are considered relevant for explaining the differences in the two groups (Eriksson et al., 2006). The validation of the PLS-DA model was checked using cross-validation and response permutation testing. Cross-validation assesses the predictive power of the model by Q2Y while the response permutation test assess the statistical significance of the estimated predictive

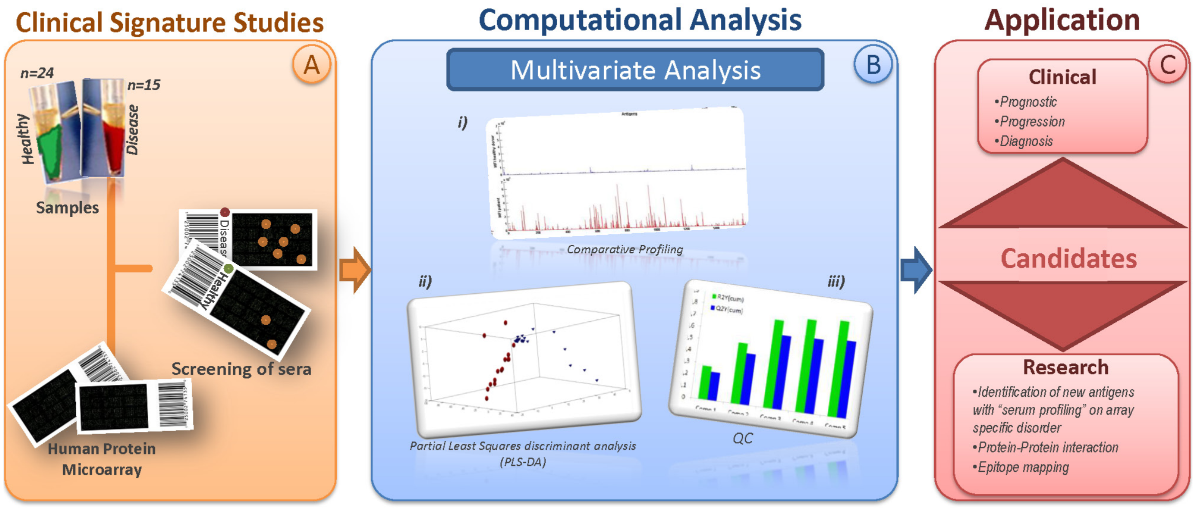

Figure 1: Schematic representation of strategy illustrating large scale serum autoantigen analysis (A) Screening of Healthy vs Patient sera by in house developed Protein microarray. (B) Multivariate Analysis: result interpretation and statistical analysis: (i) representative MFI distribuition of AIH patients (bottom panel) compared with HD (top panel) subjects, (ii) PLS-DA projection of AIH and HD samples according t1,t2 and t3 coordinates. This projection was done to identify protein profiling that distinguishes between AIH (red spheres) and HD (blue cones), (iii) plot of R2Y (explained variation) and Q2Y (predicted variation); it shows how the considered parameters change as a function of increasing model complexity. According to the cross-validation, nine components resulted significant in order to explain the relationship between the descriptor matrix and the class response; nevertheless, three components were considered to allow score plotting. (C) Investigation and further applications on the identified markers.

power and test the model for overfitting due to the chance correlation. In this test, only the class labels is randomly reordered (50 times). A model is fitted to the new Y-data and new estimates of R2Y and Q2Y values are calculated. The distribution of the R2Y and Q2Y, based on random data, are useful for appraising the validity of the model (Eriksson et al., 2006).

Results and Discussion

In order to identify a set of protein signatures linked to autoimmune liver disease, we have processed protein microarray data generated from sera of 15 AIH patients and 24 HD subjects, as shown in Figure 1, panel A. AIH sera displayed a higher reactivity toward autoantigens than HD sera as documented by the intensity of recognition signals (MFI). To detect differences between the two clinical conditions a multivariate statistical analysis was performed. As a first step, an unsupervised approach by means of PCA was applied to the full data set. This preliminary exploration by PCA was done in order to screen for outliers and to survey possible groupings, useful for efficiently directing further modeling efforts with more innovative approaches such as the PLS-DA. On the basis of the PCA score plot, a rough separation was observed owing to misclassification of one sample; thus, this sample was removed from further analyses due to its ambiguous behavior. The PLS-DA modeling has been based on the reduced data set of 38 samples described through 1296 features (X matrix). We created a dummy matrix of two Y-variables expressing diagnosis of the sera samples. Data were standardized to have mean 0 and standard deviation 1. The number of significant components was determined using cross-validation; this yielded nine components with an R2Y of 0.91 and a cross-validated R2 (Q2) of 0.75. However, a three component model was generated to enable the construction of the three dimensional score plot (Eriksson et al., 2006). There is clear discrimination between the two groups according to their clinical conditions. The model also gives the possibility to obtain a quantitative measure of the discriminating power of each autoantigens by means of VIP. X-variables characterized by VIP values larger than 1 have major importance for modeling the responses. After closer examination, autoantigens were, then, selected according to (i) VIP scores >1.0 and (ii) the recognition frequency; self proteins were regarded as potential autoantigens if they were recognized by a delta difference recognition of 25% between AIH and HD population. In this way, a final list of 27 autoantigens was generated, that allowed good discrimination of the two populations of sera. At present, the study may deepened at the biological level by further validation of these dominant features using proteomic analysis technique. Furthermore, we applied the response permutation testing to provide an estimate of the significance of a Q2Y value, we have permuted the response randomly 50 times and computed the new model with the original X-data matrix and reordered Y-data. For each derived model, both R2Y and Q2Y values were calculated and then compared with the estimates of the R2Y and Q2Y of the real model. On the basis of the validation plot, the Q2Y distribution is sign of high predictive validity of the original model indeed it is impossible to obtain a model with the same predictive value by chance.

In conclusion, we presented a multivariate approach as an effective alternative to classical univariate tools for the analysis of proteomics data for signature selection in autoimmune liver diseases. Combining multivariate modeling with protein microarray proves to be a successful tool for the discrimination of the different classes of samples and for the identification of the autoantigens responsible for class separation by means of VIP. This method could be applied for a fast screening of human protein microarrays to discriminate different clinical conditions representing a useful complementary analysis in the routine of a proteomic laboratory. However, further studies are necessary in order to extend the approach here described to different data set for verifying the chance to extrapolate and generalize classification rules.

Acknowledgements

This work was primarily supported by a Grant from Fondazione Cariplo and a grant FIRB from the Italian Ministry of University and Research (MIUR). We'd like to thank Fondazione IRCCS Ospedale Maggiore Policlinico, Milan; Policlinico Sant'Orsola, Bologna; Azienda Ospedaliera Universitaria Pisana, Pisa, for kindly providing the human sera used for the screening

References

1. Bombaci M, Grifantini R, Mora M, Reguzzi V, Petracca R et al. (2009) Protein array profiling of tic patient sera reveals a broad range and enhanced immune response against Group A Streptococcus antigens. PLoS One 4, e6332. doi:10.1371/journal.pone.0006332
2. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C et al. (2006) Multi- and megavariate data analysis. Basic Principles and Applications. Umetrics AB
3. Jain AK, Duin RPW, Mao J. (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Analysis Machine Intelligence 22, 4-37. doi: 10.1109/34.824819
4. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemomterics. Chem Intell Lab System 58, 109-130. doi:10.1016/S0169-7439(01)00155-1
5. Zingaretti C, Arigò M, Cardaci A, Moro M, Marabita F et al. (2012), Mol Cell Proteomics, Sep 20. [Epub ahead of print].