Classification of serum proteome data in familial and sporadic breast cancer

d'Acierno A[1], Garrisi VM[2], Facchiano A[1,4], De Bortoli M[3], Bongarzone I[3], Tufaro A[2], Paradiso A[2], Iannelli G[2], Tommasi S[2,4]

Motivation

Almost 20% of first diagnosed breast cancer patients with characteristics of familiarity carry mutations in BRCA1 and BRCA2 genes which often lead to an altered protein pattern. Furthermore, gene or protein alterations in the other 80% of familial breast cancers were still not define. Serum proteome is the most generally informative from a medical point of view, in particular when germinal alterations is expected. SELDI-TOF approach seemed to be a suitable approach from prognostic point of view and for therapeutic selection. ProteinChip® mass spectrometry is an innovative technology that searches the proteome, allowing the creation of a panel or profile of biomarkers. However, only the application of suitable bioinformatic tools allow to reach correct interpretation of results, both for the search of specific markers as for the classification problem in diagnostic perspective. The aim of the present study was to evidence the discriminating serum profiles in familial and sporadic breast cancer.

Methods

Blood serum from 292 people consecutively visited in the O.U. of Senology of the National Cancer Center "Giovanni Paolo II" – Bari (I) from 2004 to 2006 have been collected. Ninety-two people were classified as familial breast cancer and 100 as sporadic breast cancer patients by clinical criteria with suspicious mammogram and confirmed histological diagnosis, 100 were healthy volunteers with a negative mammogram from almost 3 years or suspicious mammogram and confirmed histological diagnosis of benign lesion. Written, informed consent was obtained from all subjects upon approval of the study by the ethic committees of the institution. The collected blood was allowed to clot at room temperature for 1 h and centrifuged at 3000 rpm for 15 min. Serum samples were stored in aliquots at -80 °C until further analysis. Samples were subjected to SELDI–TOF MS profiling using the ProteinChip Biomarker System as recommended by Bio-Rad Laboratories. Aliquot of serum (5 µl) were firstly mixed with binding buffer (PBS for IMAC30 proteinchip array and 0,1 sodium acetate pH4 for CM10 proteinchip array) and then bound (in duplicate) with a randomized chip/spot allocation scheme to IMAC-Cu and CM10 ProteinChip arrays. The energy absorbing molecule (crystallization

[1] Istituto di Scienze dell'Alimentazione, CNR, Avellino, Italy [2] Istituto Tumori "Giovanni Paolo II", Bari, Italy [3] Istituto Nazionale Tumori, Milano, Italy [4] RNBIO - Rete Nazionale di Bioinformatica Oncologica

matrix) sinapinic acid was dissolved in 50% acetonitrile/0.5% trifluoroacetic acid and was applied air dried and reapplied. Spotted arrays were read using the Protein-Chip© reader PCS 4000 Enterprise (Bio-Rad). For each experimental condition, arrays were read at setting optimized for low molecular weight (2–30KDa). Data acquired from spectra as m/z and peaks intensities finally were used for the classification study. Classification is based on support vector machines with RBF kernel and a classical PCA has been used to reduce the data dimensionality; we tested our systems for different values of the scaling factor used in kernel functions and for different values of the overall energy of data (i.e. the number of components). We implemented a 10-folds cross validation that has been repeated 10 times (so deriving 100 runs for each couple of parameters). At each run, correct classification rates (CR) on the test sets have been measured that are last averaged over the 100 runs. The available data refer to samples from three classes (familial cancer, sporadic cancer, and control individuals); the classification problem has been formalised as a binary problem and so we tried to distinguish between familial and sporadic cancer, and also to distinguish each of these two classes from the control samples.

Results

The application of the classification strategy to the ProteinChip arrays results performs differently for the two chips. The analysis of CM10 results gave a best CR between familial or sporadic cancer with 68.5% of success. When applied to distinguish between familial cancer and control samples, the best CR obtained is 70.1%, and between sporadic cancer and control samples, 70.6% of success. Better results were obtained by analyzing the IMAC-Cu Protein Chip data. The success in classifying data from familial or sporadic cancer was 70.7%, while familial cancer and control samples were correctly classified with 80.3% of success, and between sporadic cancer and control samples, with 78.2% of success. These results could be further improved by fine tuning of the some strategy parameters, although the expectable increase of success is of few percentage units. Further investigations are planned to improve the understanding of the results and the CR, from both the bionformatic and experimental side of the study, and to verify some hypotheses about the reasons which determinate the different results obtained in the analysis of the data sets from the two different ProteinChips.

Contact e-mail

angelo.facchiano@isa.cnr.it