Use of powered supervised learning statistics for the uninvasive diagnosis of Renal Cell Carcinoma through urine proteome analysis

Maglietta R[1], Papale M[2], D'Onofrio V[2], Stifanelli P[1], Ranieri E[3], Battaglia M[4], Carrieri G[5], Ancona N[1]

Motivation

Renal Cell Carcinoma (RCC), the most prevalent form of Renal Cancer (RC), is often casually diagnosed in many patients due to its asymptomatic nature. Proteomic analysis of urine samples from both RCC and control patients (CTRL) would show complex protein datasets closely associated to the diseases and potentially useful for the selective diagnosis of RCC in asymptomatic patients. In the present work we aimed to evaluate the power of urine proteome analysis by means of supervised learning statistics in view of set up new and uninvasive tools for ameliorating RCC diagnosis.

Methods

Protein profiles produced by SELDI-TOF/MS analysis of urine from RCC and CTRLs were analysed by univariate statistical approaches to detected differently expressed mass peaks among sample groups. To this end, we evaluated the Wilcoxon rank sum statistic and used it to assign a statistical significance (P-value) to each mass peak. Bonferroni's correction and False Discovery Rate (FDR) were assessed to take into account problems related to multiple hypothesis testing. Successively, the data were analyzed by multivariate predictive models to discriminate RCC patients from CTRLs by using their protein profiles. To this end, we used Regularized Least Square (RLS) classifiers which exhibit low error rate on unseen subjects. The number of data used to build the model was assessed by using Cross Validation procedure varying the number of training examples. The statistical significance of the model was evaluated by using nonparametric permutation tests. To assess the accuracy of the model on unseen subjects, we developed a methodology aiming to estimate the class label of a new subject belonging to a separate validation data set; it is important to emphasize that the procedure is unbiased because the examples in the validation test were never employed in the previous analysis. In particular, we randomly extracted n subjects from the original data set to train the RLS classifier; successively, we tested the

[1] Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR, Via Amendola 122/D-I, 70126, Bari [2] Proteomic core Facility Research Center Bioagromed, University of Foggia, Via Napoli 52, 71122, Foggia [3] Clinical Pathology, Faculty of Medicine, University of Foggia, Viale L. Pinto 1, 71122, Foggia [4] Urology, Faculty of Medicine, University of Bari [5] Urology, Faculty of Medicine, University of Foggia

classifier on each profile in the validation test. The procedure was repeated " times, and the prediction error e on each new subject was given by the mean of the errors in the " models. The statistical significance of e was evaluated by a nonparametric permutation test.

Results

In this study, 102 protein expression levels were produced by SELDI-TOF/MS analysis of urine from 20 samples of RCC and 15 samples of CTRL. Wilcoxon test detected 48 proteins (P-value < 0.05) and 12 showed a P-value lower than the cut-off adjusted by the Bonferroni correction (P-value<0.0005) with FDR < 0.001: C06516_4, C05569_9, C03086_5, C04136_3, C03891_1, C06531_5, C07049_8, C03346_2, C05084_9, C05353_5, C03032_8, C08017_6. The multivariate statistical analysis was carried out by using RLS classifiers with different training set size. We found that the optimal training set size was n=15 having the lowest error rate (e = 18.4% and p-value = 0.026). Finally, a validation set composed by 17 new samples of RCC and 6 new samples of CTRL was used to test and validate our predictive model. We used the described methodology to evaluate the error rate e and its p-value on each patient in the validation set setting by using " = 5000. The results are illustrated in the table and figure. Only two RCC samples, patient 12 and patient 13, are misclassified showing an error rate greater than 50%. The labels of the remaining 21 examples were correctly assigned by our method. In particular, 15 out of the 23 examined examples showed a prediction error e lower than the CV_error with a p-value < 0.05, highlighting the excellent performances of our method. Moreover, the mean error on the validation samples was 16.9% perfectly comparable to the CV_error. sample e(%) p-value 13 77 0.978 12 69 0.948 16 49 0.895 2 47 0.67 8 41 0.523 7 34 0.57 4 31 0.472 1 17 0.208 14 9 0.11 18 9 0.04 21 1 0.002 10 1 0.002 17 1 0.003 19 0 0.001 3 0 0 6 0 0 9 0 0 5 0 0 23 0 0 22 0 0.002 11 0 0.001 15 0 0 20 0 0

Contact e-mail

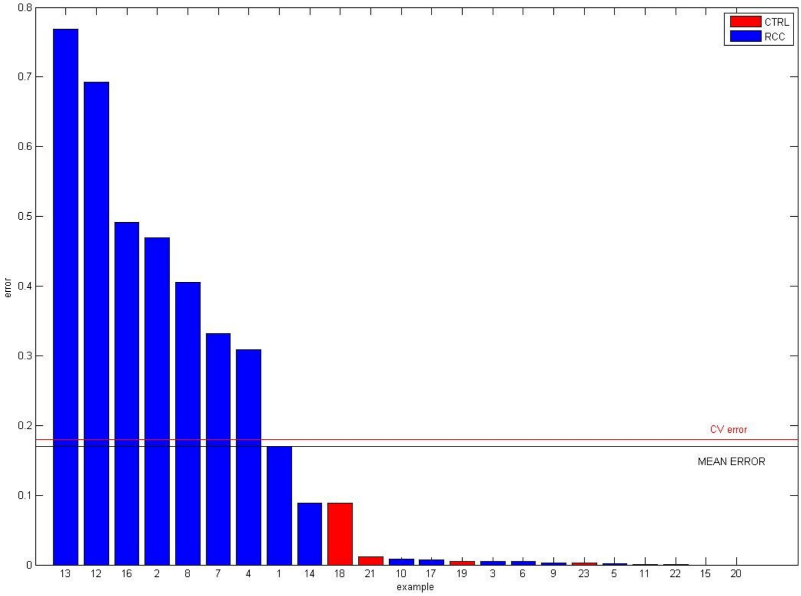
ancona@ba.issia.cnr.it

Supplementary information

Funding

This work was supported by grants from Regione Puglia, Progetto Strategico PS_012 and Progetto Reti di Laboratori Pubblici di Ricerca BISIMANE.

Image



CTRL
RCC
CV error
MEAN ERROR
error
example
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
13
12
16
2
8
7
4
1
14
18
21
10
17
19
3
6
9
23
5
11
22
15
20