Serum seleno-proteins status for colorectal cancer screening explored by data mining techniques - a multidisciplinary pilot study

Marco Roman a,⁎, Petru Jitaru b, Marco Agostini c, Giulio Cozzi a, Salvatore Pucciarelli c, Donato Nitti c, Chiara Bedin d, Carlo Barbante a,e,f

a Institute for the Dynamics of Environmental Processes (IDPA-CNR), Dorsoduro 2137, 30123 Venice, Italy
b Institut Polytechnique LaSalle Beauvais, Département des Sciences et Techniques Agro-Industrielles, 19 rue Pierre Wagué BP 30313, 60026 Beauvais, France
c University of Padova, Department of Oncologic and Surgical Sciences, Via Gattamelata 64, 35100 Padova, Italy
d Institute of Molecular Sciences and Technologies (ISTM-CNR), Corso Stati Uniti 4, 35127 Padova, Italy
e University Ca’ Foscari of Venice, Department of Environmental Sciences, Dorsoduro 2137, 30123 Venice, Italy
f Accademia Nazionale dei Lincei, Centro B. Segre, via della Lungara 10, 00165 Roma, Italy

A R T I C L E   I N F O

Article history:
Received 30 September 2011
Received in revised form 31 January 2012
Accepted 12 February 2012
Available online xxxx

Keywords:
Colorectal cancer
Seleno-proteins
Inductively coupled plasma-mass spectrometry
Logistic regression analysis
Classification trees
Artificial neural networks

A B S T R A C T

In this case–control pilot study a recent method based on HPLC hyphenated to ICP-MS was employed for the quantification of serum glutathione peroxidase type 3 (GPx3), seleno-protein P (SelP) and seleno-albumin (SeAlb) in 42 patients with colorectal cancer (CRC) and 20 controls. Patients with early cancer stage (TNM I) showed a significantly higher level of SeAlb (19±3 ng/mL) in respect to both metastatic CRC patients (TNM IV, 16±4 ng/mL) and healthy controls (16±3 ng/mL). Classification models based on logistic regression analysis, classification trees and artificial neural networks were constructed using seleno-proteins concentrations as predictors. Neural networks lead to the best performances, up to 95% of corrected predictions in TNM I vs. controls discrimination. These results suggest a potential association between individual seleno-proteins and CRC progression. Age and radiochemotherapy were assessed as confounding factors, showing no significant effects. Still, SeAlb level tended to reduce with the age in healthy persons, but did not in CRC patients. Seleno-proteins concentration was also compared with a number of clinical parameters considered as prognostic factors in CRC. Significant Spearman’s correlations were revealed between SeAlb and SelP, and presence of peritumoral lymphocytic infiltration (ρ = −0.57 and ρ = −0.37, respectively); and SeAlb and degree of cellular differentiation (grading, ρ = −0.37). This study marks the importance to systematically introduce speciation analysis and multidisciplinary approaches in the investigation of the role of seleno-proteins as a potential combined biomarker for CRC.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Selenium (Se) is an essential trace element whose important function in human health has attracted great attention in the last decade. Its role in oxidative stress regulation makes it potentially involved in the development, progression and prognosis of several diseases, in particular cancers [1]. Colorectal cancer (CRC) is the third most commonly diagnosed form of cancer in women and fourth in men worldwide [2], and its association with Se has been investigated in a number of studies. Some of them observed a significantly lower level of Se in serum of CRC patients than in healthy subjects, or a significantly higher risk to develop CRC and lower cumulated cancer-related survival rates for subject presenting low serum Se levels [3–5]. A similar behaviour was noticed in patients affected by colorectal adenoma, the precursor lesion in most CRC cases [6–9]. Selenium levels have also been associated with the CRC stage (from adenomatous polyps to local and metastatic cancer, respectively) [10]. However, other studies regarding Se and colorectal adenoma did not observe significant correlations [11–13]. These contradictions reflect a more general picture emerging from a wide number of trials and cross-sectional studies on the cancer preventive action of supplemented Se, that observed both beneficial than irrelevant effects, suggesting a possible pathogenic specificity [14–17]. Hence, the association between Se status and cancer is probably more complex than initially expected.

A critical aspect is that different biological functions of Se are carried out by individual proteins (Se-proteins), where it participates to the active site. Se is incorporated into Se-proteins as the amino acid Se cysteine (SeCys) through a genetically encoded pathway. Complex regulatory mechanisms showed to be responsible for the preferential incorporation of Se into certain Se-proteins, so that in marginal element deficiency the metabolism of these compounds is unaffected whereas the content of other species might be decreased [18,19]. Se-proteins have to be also distinguished from the Se-containing proteins, that are biologically inactive (from the point of view of Se)

0026-265X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.microc.2012.02.004
were suggested as alternative biomarker for Se status assessment, including for cancer studies, which could reflect more accurately the real dynamics of the element [12,22]. In this context, the human plasma/serum Se-proteins glutathione peroxidase type 3 (GPx3) and Se-protein P (SelP), as well as the Se-containing protein Se-albumin (SeAlb, for simplicity also referred below as a Se-protein) are of high interest. An inverse correlation has been observed between various types of cancer (including CRC) and serum GPx3 or plasma SelP [4,23,24], but other works did not confirm this data [12,25]. To our knowledge only a single study by Early et al. [12] took into account the complete plasma Se-proteins status in CRC. In such work GPx3, SelP and total Se were determined in distinct steps recurring to enzymatic assay, radioimmunoassay and fluorometry, respectively, while the SeAlb level was calculated by difference, and only basic data analysis was provided. Thus, new investigations are needed to elucidate the role of serum Se-proteins in CRC on the bases of both robust analytical methods and comprehensive data elaboration.

For this type of applications, predictive data mining techniques has been increasingly used in recent years [26]. These approaches are able to combine a wide range of patients-specific different information to obtain descriptive models that can turn in decision systems effectively supporting clinical processes such as diagnosis, prognosis and treatment. Data mining for classification proposes manipulates data sets to elucidate hidden patterns by a variety of methods. Among them artificial neural networks [27,28], k-nearest neighbours [29], hierarchical clustering [30] and principal component analysis [31] have been already explored to discover unknown data structures in proteomic and genomic profiling of CRC, allowing not only to discriminate CRC from normal tissues but also to properly classify tumor subgroups. Support vector machine has been also applied to serum proteomic profiling to predict the early-stage response of CRC to radiochemotherapy [32]. Other data mining methods among the most used for clinical tasks include decision trees and rules, logistic regression, and the more classical chemometric techniques [26,33].

The aim of our study was to combine a newly developed analytical methodology for the simple, fast and accurate simultaneous determination of GPx3, SelP and SeAlb in human serum [34] with clinical information and data mining tools. Samples from 42 patients affected by CRC and 20 healthy control subjects were analyzed in this pilot work, in order to assess the possible association between Se-proteins pattern and the presence of cancer, progression stage and also prognostic criteria. A comprehensive post-hoc statistical approach based on data mining techniques including logistic regression analysis, classification trees and artificial neural networks was addressed in order to explore the potentialities of Se-proteins level as an integrated biomarker for CRC screening.

2. Materials and methods

2.1. Patients information and samples collection

The study group consisted of 42 patients affected by CRC that underwent a colonoscopy control for colic disorder and had a histologically proven colorectal adenocarcinoma. Diagnosis was performed at the University Hospital of Padova (Italy), then patients underwent a surgical intervention in the same hospital at the Department of Surgical Clinic II. In order to avoid overlapping of the pathologic stages we selected 22 patients at stage I (infiltration of bowel wall without lymph nodes or distant metastases) and 20 patients at stage IV/metastatic cancer (presence of distant metastases) according to the tumor node metastasis (TNM) classification of the American Joint Committee on Cancer (AJCC) [35]. The control group included 20 healthy (H) subjects who did not present genetic syndrome and underwent colonoscopy control in the same hospital for colic disorders (haemorrhoids, diverticulosis or functional disorders), but resulted negative for CRC as well as inflammatory pathologies.

Blood was collected just before surgery (CRC patients) or after colonoscopy (H subjects). The serum was obtained by centrifugation of the blood at 3000 rev/min for 10 min, divided into aliquots of 1.5 mL and stored at −20 °C until analysis. A human serum certified reference material (CRM) BCR-637 (certified for total Se content) from the Institute for Reference Materials and Measurements (IRMM, Geel, Belgium) was used throughout for quality control.

Informed consent was obtained and the study was performed in conformance with the Declaration of Helsinki ethical guidelines.

2.2. Clinical parameters

A set of clinical information, reported below and resumed in Table 1, was also collected for CRC patients. Typical histological and tissue-based prognostic/predictive markers [36] were selected to assess their possible association with serum Se-proteins level.

2.2.1. Histological markers

Histopathological studies were carried out on the surgically removed tissues. Histotype was defined according to the presence of mucinous histological cancer subtype [37]. Presence of peritumoral lymphocytic infiltration was assessed according to the Jass’ classification system [38]. Vascular invasion was defined as the presence of malignant cells within endothelial cell-lined blood vessels beyond the muscularis...
propría [39]. Tumor grading was assessed according to the three-grade system of WHO Classification: well differentiated (I), moderately differentiated (II) or poorly differentiated (III) [35]. Tumor staging according to the TNM classification (as reported above) was also histologically assessed.

2.2.2. Tissue-based markers

Tissue-based studies were carried out on the extracted DNA by PCR and gel electrophoresis searching for specific markers in order to assess microsatellite instability (MSI) [40] and loss of heterozygosity (LOH) [41]. Tumors were classified as MSI-H if ≥30% markers demonstrated instability, MSI-L if <30% demonstrated instability, and MSS if no marker exhibited instability. Overall 18q LOH positivity was defined as ≥1 informative markers with LOH and 18q LOH negativity as ≥2 informative markers and no evidence of LOH.

Maximum tumor diameter was measured by magnetic resonance imaging (MRI).

2.2.3. Potential confounding factors

Family history was considered for both H and CRC subjects up to second degree family members. The parameter defines the belonging of patients to the following risk groups: sporadic if no CRC cases were registered; suspected hereditary nonpolyposis colorectal cancer (s-HNPPC) if at least one of the Revised Bethesda Criteria was satisfied [42]; familial adenomatous polyposis (FAP) if corresponding cases were registered [43]. Preoperative treatment consider if patients underwent radiochemotherapy (RCHT) before the surgery. Tumor site consider if cancer was present in colon or rectum. Age, gender, actual status (dead or alive) and cancer recurrence after radical surgery were also considered in the pool of clinical information.

2.3. Analytical procedure for Se-proteins determination and quality control

Determination of GPx3, SelP and SeAlb (in terms of Se content) in human serum was achieved by using tandem ion exchange (AE) and double affinity (AF) high performances liquid chromatography (HPLC) separation coupled on-line with inductively coupled plasma-mass spectrometry (ICP-MS). Details regarding this method, the standards/reagents and the optimum parameters were described in a previous study [34].

The quality control for Se-proteins quantification was ensured by the analysis of the BCR-637 CRM (certified total Se level=81±7 ng/mL). The levels of GPx3, SelP and SeAlb in BCR-637 were determined in a previous work [34] by using the same method as employed in this study and hence its use for quality control can be considered satisfactory. The levels of Se-proteins in BCR-637 obtained as described above are: 13±1 ng/mL for GPx3, 59±3 ng/mL for SelP and 16±2 ng/mL for SeAlb. The BCR-637 serum was analyzed every 10 samples and the average value (n=5) obtained for Se-proteins was 12±3 ng/mL for GPx3, 55±4 ng/mL for SelP and 16±2 ng/mL for SeAlb. These values and the reference levels were compared by t-test and by calculation of difference between averages in percentage, confirming the accuracy of the analysis in this study (p for t-test <0.05, difference between averages <10%). The sum GPx3+SelP+SeAlb in BCR-637 (83±5 ng/mL) was also in good agreement with the certified total Se (81±7 ng/mL), hence the sum of the three concentrations was considered to approximate accurately the total Se level.

Due to the low available volume of sample, no replicate measurements were possible. To assess the overall weight of intersample uncertainty, method reliability was estimated as follows. The precisions of the method (for individual Se-proteins) as reported in the reference paper [34] were assumed to be representative of the intersample RSDs. The total RSDs were calculated for the each group of patients (H, CRC, TNM I, TNM IV) and the corresponding intersample RSDs were calculated by subtracting the intrasample RSDs from total RSDs. The ratio intersample RSD/total RSD (intragroup correlation coefficient, ICC) resulted in all cases >0.78. This indicates good reliability of the method and implies that intrasample uncertainty can be considered negligible in respect to the intersamples variability within each group of patients.

2.4. Statistical data treatment

Mean values, SD and principal distribution parameters of Se-proteins level were calculated for CRC, TNM I, TNM IV and H groups. The normality of their distribution was tested by Shapiro–Wilk (S-W) test [44], in order to correctly apply the t-test for paired groups comparison, and if necessary the data were treated by Box–Cox transformation or the non-parametric Kolmogorov–Smirnov (K-S) test [44] was applied instead of the t-test. Se-proteins level was also tested for possible confounding effects by analysis of covariance (ANCOVA). When necessary, data were adjusted by removing the component explained by the linear regression model.

In order to deeply explore the potential of Se-proteins as markers of CRC or staging classification models were constructed using the three concentrations as input variables and adopting a number of alternative approaches. Logistic regression analysis (LRA) was applied as it is the most used method in case–control epidemiologic studies. Classification trees (CTs) were developed as they consist in the simplest strategy for decision making, and are widely used in diagnostics as well. Backpropagation and probabilistic artificial neural networks (BpANNs and PrANNs, respectively) were also applied. Neural networks are powerful classifiers, but still little used in clinical studies due to the difficult handling for non statisticians and to the less accessible interpretation of results. Even if all methods could be applied for multiple (>2) groups comparison, we decided to develop independent models for all possible couples of groups (CRC vs. H; TNM I vs. H; TNM IV vs. H and TNM I vs. TNM IV) in order to extract a more complete information on specific classification potentials.

Statistica 10.0 (StatSoft Inc., USA), Origin 7.5 (OriginLab Corp., USA) and Office Excel 2003 were used for data analysis and graphical elaboration.

2.4.1. Logistic regression analysis

In LRA a linear model is constructed to represent a categorical response variable as a function of one or several independent continuous variables [45]. For our diagnostic purposes, the response of the model is the logit transformation of a probability p of belonging to a group (vs. another group), while the input variables are the Se-proteins concentrations. The basic structure of the model is:

\[
\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{GPx3} + \beta_2 \text{SelP} + \beta_3 \text{SeAlb}
\]

2.4.2. Classification trees

Classification trees are recursive partitioning classifiers [46]. The model is structured in nodes that form a tree-like diagram. The data set is presented at the base of the tree and iteratively split at each node according to a threshold value applied to one of the predictor

Please cite this article as: M. Roman, et al., Serum seleno-proteins status for colorectal cancer screening explored by data mining techniques - a multidisciplinary pilot study, Microchem. J. (2012), doi:10.1016/j.microc.2012.02.004
variables, i.e. the Se-proteins levels. The best threshold and ramification levels are calculated by a specific discrete function. In this study we adopted the C4.5 algorithm [47], which comprises two steps: induction based on gain ratio as splitting criteria and pruning based on the misclassification error. The pruning step is necessary to avoid overfitting of the data. A group is assigned at each subset of data when it reaches the termination of a branch (leaf), thus CTS directly provide an assignment rather than a probability. Goodness of fit was assessed by Gini index. The validity of each CT model was tested by a 3-fold leave-10%-out cross-validation.

2.4.3. Artificial neural networks

Artificial neural networks are computational models constructed to imitate the structure and behaviour of neurons in the human brain [46]. A number of interconnected processing units (neurons) can be trained for pattern recognition through processing in parallel an input set to generate the appropriate output. The network topology comprises an input layer with 3 neurons corresponding to the observed variables (Se-proteins levels), and two active layers: a hidden layer with a variable number of neurons (n), and an output layer with 2 neurons, one for each response code (CRC, TNM I, TNM IV or H). Thus, the networks topology is described by the code 3:n:2, as represented in Fig. 1. Similarly to LR models, the output of ANNs is a probability to which a threshold value is applied for deriving the final decision.

Multi-layer perceptron (MLP) ANNs are based on neurons which compute a non-linear function of the scalar product of the input vector and a weight vector [49]. When based on the back-propagation algorithm (BrANNs), an iterative training is performed by correcting each actual weight with a term directly proportional to the input and to the error, according to the theory of gradient descent learning. In probabilistic ANNs (PrANNs) the hidden neuron activation is determined by a radial basis function (RBF) which computes the distance between the input vector and a prototype vector [49].

Both BrANN and PrANN models were developed in this work using the standardized concentrations as the input set. Training sets were randomly selected to represent 70% of each group's data. The remaining data set was divided into two equal parts to be adopted as test set and validation set, respectively. An initial raw optimization of the network was performed by exploring different topologies with n between 3 and 10 for BrANNs or between 10 and 14 for PrANNs. Cross entropy was adopted as error function, while identity, logistic, tanh and exponential functions were tested for hidden neurons activation. Ten thousand networks were randomly generated, and the best 100 were retained. A screening of the performance allowed to select the optimal features in terms of topology and transfer function for the hidden neurons. Five thousand new networks were then trained following the conditions and topology of choice. The best network was retained and evaluated.

2.4.4. Performance of the classification tools

The performance of the classification models was assessed in terms of sensitivity (SEN), specificity (SPE) and ratio of corrected predictions (RCP), as in Eqs. (2)-(4).

\[ \text{SEN} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \]  
(2)

\[ \text{SPE} = \frac{\text{true negative}}{\text{true negative} + \text{false positive}} \]  
(3)

\[ \text{RCP} = \frac{\text{true positive} + \text{true negative}}{n \text{ cases}} \]  
(4)

The receiving operating characteristics (ROC) curves [50] were also represented for each model. In ROC graphs SEN is plotted on the y-axis and 1-SPE is plotted on the x-axis. In the ROC space, the performance of a classifier is described as a curve calculated by varying the threshold value to be applied to the output probability (for LR and ANN models) or by changing the class assignment at each leaf (for CTS). The diagonal line y=x in the ROC graph represents the strategy of randomly guessing a class. The performance of a model is higher as much its curve drift away from the bisector in the upper part of the graph, i.e. the corresponding area under the curve (AUC) approaches 1. In this study the AUC of the best models was calculated and tested for significance in respect to that of the bisector (0.5) by Hanley & McNeil method [51]. If significant, all couples of ROC curves for each classification goal were tested among themselves for significant difference in order to directly compare the models.

2.4.5. Association with prognostic criteria

The possible association between Se-proteins level in CRC patients and categorical prognostic criteria was tested by calculation of the Spearman’s rho correlation coefficients [44]. Based on correlations, the more interesting prognostic criteria were selected as new grouping variables to compare Se-proteins concentration via S-W and t-tests.

Table 2

<table>
<thead>
<tr>
<th>Protein</th>
<th>H (n=20)</th>
<th>CRC (n=42)</th>
<th>TNM I (n=22)</th>
<th>TNM IV (n=20)</th>
<th>H vs. CRC</th>
<th>H vs. TNM I</th>
<th>H vs. TNM IV</th>
<th>TNM I vs. TNM IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPx3</td>
<td>18 ± 4</td>
<td>17 ± 3</td>
<td>17 ± 3</td>
<td>17 ± 4</td>
<td>0.46</td>
<td>0.67</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.66)</td>
<td>(0.62)</td>
<td>(0.96)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SeIP</td>
<td>59 ± 10</td>
<td>62 ± 13</td>
<td>65 ± 13</td>
<td>59 ± 12</td>
<td>0.42</td>
<td>0.91</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.09)</td>
<td>(0.12)</td>
<td>(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SeAlb</td>
<td>16 ± 3</td>
<td>18 ± 4</td>
<td>19 ± 3</td>
<td>16 ± 4</td>
<td>0.13</td>
<td>&lt;0.01</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.42)</td>
<td>(&lt;0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Se</td>
<td>93 ± 16</td>
<td>97 ± 16</td>
<td>101 ± 17</td>
<td>92 ± 15</td>
<td>0.46</td>
<td>0.17</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.23)</td>
<td>(0.27)</td>
<td>(0.02)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please cite this article as: M. Roman, et al., Serum seleno-proteins status for colorectal cancer screening explored by data mining techniques - a multidisciplinary pilot study, Microchem. J. (2012). doi:10.1016/j.microc.2012.02.004
3. Results

3.1. Comparison between CRC patients and healthy subjects based on their Se-proteins status

GPx3, SelP and SeAlb mean concentrations, SD and p of the t-test are reported in Table 2, and the corresponding box-plots are shown in Fig. 2. The values are compatible with the literature [12,52]. No significant differences were observed in the level of GPx3, SelP, SeAlb, nor in their sum, between CRC and H groups. We noticed just a slightly lower GPx3 mean level and higher SelP, SeAlb and total Se mean levels in CRC samples, although these differences did not reach statistical significance. Rather, significantly higher SeAlb concentration was observed for TNM stage I patients in comparison with H (p < 0.01) and TNM IV (p < 0.05) subjects.

As reported by Chan et al. [53], the long-course preoperative treatment by radiochemotherapy (RCHT) may alter the TNM classification by downstaging of T and N parameters. Nine patients in our study underwent preoperative treatment (see Table 1), therefore the possible confounding effect of this prognostic criterion was also evaluated by calculating the adjusted p, reported in brackets in Table 2. When adjusted for RCHT, TNM I and IV stages were significantly different not only in terms of SeAlb level (p < 0.01) but also for SelP (p < 0.05) and total Se (p < 0.05), while GPx3 remained not significant (p > 0.05). Thus, individual Se-proteins level and total Se were not significantly different between RCHT treated and untreated patients (p > 0.05 in all cases); RCHT treated patients were not significantly concentrated in one of the two TNM groups (4 RCHT treated patients were in TNM I stage and 5 were in TNM IV stage), but it was found that RCHT interacts with the TNM stage enough to increase the SelP, SeAlb and total Se level significance by TNM stage. In order to correct the comparison between healthy subjects and TNM I-IV groups, patients with no information regarding their RCHT treatment (10 from TNM I group and 4 from TNM IV group) were removed from the data set, and for the remaining patients GPx3, SelP and SeAlb levels were adjusted for RCHT before the application of the t-test. The adjusted ps (see Table 2) were in some cases highly different from the unadjusted data set, but no changes in significance were observed: only SeAlb concentration was significantly different between H subjects and TNM I patients. Of course, the further reduction of sample size may affect this result.

Age (classified in thirds) and gender were also tested as confounding factors, but no significant effects were observed. However, SeAlb level decreased slightly with the age H subjects but not in CRC patients. The same behaviour was seen also by calculating the skewness of the ratio Age/SeAlb in different groups: 0.57 for CRC, 0.23 for TNM I, 0.66 for TNM IV and 0.86 for H subjects. All groups were right-skewed (skewness > 0), but in H subjects the skewness was higher in respect to CRC patients, in particular the TNM I group, therefore they tended to present lower dispersion in higher Age/SeAlb ratios. The kurtosis of the ratio Age/SeAlb was 0.70 for CRC, 0.61 for TNM I, 0.21 for TNM IV and 1.82 for H. The higher value of kurtosis in H subjects confirmed the tendency of the ratio to be more concentrated around the median.

3.2. Classification

The overall performances of classification models are summarized in Table 3. If their ROC curve was significant in respect to the bisector the best model for each technique was selected and reported to be compared with the others. Considering the whole grouping combinations the significant LRA, CT, BpANN and PrANN models had AUC in the range 0.71–0.99.

3.2.1. Logistic regression analysis

Different input sets were tested by combining GPx3, SelP, SeAlb concentration, the ratio of GPx3/SeAlb and the product SelP·SeAlb to assess their possible confounding and interacting effects. In models construction GPx3, SelP and SeAlb levels were tested individually and subsequently combined. GPx3 and SeAlb were significant (p < 0.05) in all models, and were confounding (if one was missing the other was not significant) as well as interacting (the introduction of their product produced a synergetic effect, in this case dominant with respect to the confounding effect). For discrimination between CRC vs. H groups SelP as individual input variable and also combined with the other proteins was not significant (p > 0.05). This implies that it could be uncorrelated with the presence of cancer, and did not capture the effect of the other variables. The model showed an almost null classificatory power due to the extremely low SPE (10%). In the cases of TNM I/IV vs. H discriminations SelP was significant (p < 0.05), but the introduction of the product SelP·SeAlb gave inhibitory effects. The performance for TNM I vs. H groups classification was notably higher as SPE (90%), even if not with optimal SEN (82%). The TNM I group was also slightly differentiable from the TNM IV group, but with worst performance.

3.2.2. Classification trees

To avoid overfitting problems, classification trees needs to be too simple, particularly in respect to the low number of input variables, to guarantee enough classification power. Only for TNM I vs. H classification a significant model was obtained, represented in Fig. 3. The principal discriminating variables were SeAlb (first splitting) and GPx3 (third splitting), in agreement with the previous observation that TNM I patients are characterized by high SeAlb and low GPx3 levels.

3.2.3. Artificial neural networks

Significantly classifiers were constructed with BpANN for all comparisons with the exception of CRC vs. H. The best performance (~90%
for both SEN and SPE) was obtained for TNM I vs. H, slightly higher than for TNM IV vs. TNM I. Fair performances, with SEN and SPE in the range of 80%, were obtained also for the TNM IV vs. H classification. In all cases the topology 3:3:2 and a logistic transfer function were selected as optimal features. PrANNs were also a feasible tool for all the considered classifications, with best performances for TNM I vs. H (95% for both SEN and SPE). For PrANNs the topology 3:10:2 was selected in all cases with the exception of TNM I vs. H classification (3:14:2 topology).

3.2.4. Comparison of classification tools

The ROC curves of the best models, represented in Fig. 4, graphically confirm the results reported in Table 3. For CRC vs. H classification the PrANN appears to be the only model showing a ROC curve that qualitatively differs from the bisector. All the other classifications resulted well predictable also by BpANNs and LR models. PrANNs and BpANNs tend to perform equally and better than LR models, that in turn outperforms the CT obtained for the TNM I vs. H classification. Such hierarchies are uniformly respected across all the prediction thresholds, with the only exception of PrANNs which performed slightly better than BpANNs only in the lower part of the curve. Classifications involving the TNM I group were predicted with the best general accuracy, in particular vs. the H group. According with the Hanley & McNeil test for the AUCs, reported in Table 4, the performances of LRA and BpANN, LRA and PrANN, CT and PrANN were statistically different ($p<0.05$) in the TNM I vs. H comparison, while the performances of LRA and BpANN were statistically different for TNM IV vs. TNM I. ANNs were confirmed to be the best performing models for our application. PrANNs in general gave slight worst performance in comparison to BpANNs, with the exception of TNM I vs. H discrimination, where they resulted in the best absolute performance.

3.3. Association between Se-proteins level and prognostic criteria in CRC patients

3.3.1. Correlations

GPx3 was not significantly correlated with any of the criteria considered in this study. SelP showed a significant negative correlation

<table>
<thead>
<tr>
<th>Groups</th>
<th>Models</th>
<th>Difference between areas (CL)</th>
<th>z-statistic</th>
<th>p-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRC vs. H</td>
<td>LRA vs. PrANN</td>
<td>0.09 (-0.07 to 0.26)</td>
<td>1.10</td>
<td>0.27</td>
</tr>
<tr>
<td>TNM I vs. H</td>
<td>LRA vs. CT</td>
<td>0.01 (-0.15 to 0.16)</td>
<td>0.12</td>
<td>0.91</td>
</tr>
<tr>
<td>LRA vs. BpANN</td>
<td>0.08 (-0.04 to 0.19)</td>
<td>1.32</td>
<td>0.01*</td>
<td></td>
</tr>
<tr>
<td>LRA vs. PrANN</td>
<td>0.15 (0.04 to 0.26)</td>
<td>2.63</td>
<td>&lt;0.01*</td>
<td></td>
</tr>
<tr>
<td>CR vs. BpANN</td>
<td>0.07 (-0.06 to 0.20)</td>
<td>1.04</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>CR vs. PrANN</td>
<td>0.14 (0.02 to 0.26)</td>
<td>2.37</td>
<td>0.02*</td>
<td></td>
</tr>
<tr>
<td>BpANN vs. PrANN</td>
<td>0.07 (-0.01 to 0.16)</td>
<td>1.67</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>TNM IV vs. H</td>
<td>BpANN vs. PrANN</td>
<td>0.01 (-0.13 to 0.14)</td>
<td>0.11</td>
<td>0.91</td>
</tr>
<tr>
<td>TNM I vs. H</td>
<td>LRA vs. BpANN</td>
<td>0.17 (0.02 to 0.33)</td>
<td>2.17</td>
<td>0.03*</td>
</tr>
<tr>
<td>TNM IV</td>
<td>LRA vs. PrANN</td>
<td>0.13 (-0.05 to 0.30)</td>
<td>1.42</td>
<td>0.16</td>
</tr>
<tr>
<td>BpANN vs. PrANN</td>
<td>0.05 (-0.11 to 0.20)</td>
<td>0.58</td>
<td>0.56</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Best decision tree for classification between TNM I and H groups. The number of subjects belonging to each group (following the same order) is indicated, as well as the final classification result.

Fig. 4. ROC curves of the best classification models. The continuous gray line is the reference bisector.
with the presence of peritumoral lymphocytic infiltration ($p = -0.57$, $p < 0.01$) and interestingly, although not statistically significant, it was correlated with loss of heterozygosity (LOH; $p = -0.33$, $p = 0.05$). SeAlb was significantly correlated with infiltration ($p = -0.37$, $p = 0.05$) and degree of cellular differentiation (grading; $p = -0.37$, $p = 0.05$).

### 3.3.2. Groups comparison

On the basis of correlations, tumour infiltration, grading and LOH were selected as grouping variables for the comparison of Se-proteins by S-W and subsequent t-test. All data were normally distributed (S-W test: $p > 0.05$). Significantly ($p < 0.05$) higher levels of SeIP were found in patients without infiltration (69 ± 9 ng/mL) in comparison with those with infiltration (57 ± 12 ng/mL); the same result was obtained for patients without LOH (64 ± 12 ng/mL) in comparison with patients who presented it (55 ± 12 ng/mL). SeAlb was assessed with significantly ($p < 0.05$) higher concentrations in patients without infiltration (19 ± 3 ng/mL) in comparison with patients who had it (17 ± 3 ng/mL) and for patients with grading I (19 ± 2 ng/mL) in comparison with patients with grading > I (17 ± 3 ng/mL).

### 4. Discussion

Faecal occult stool testing and colonoscopy are the currently recommended screening tests for CRC, but only 35% of adults follow these guidelines, and there are still doubts regarding their cost-effectiveness balance and patients compliance [54]. In order to overcome these drawbacks, less invasive screening tests for the early detection of CRC have been studied on the bases of identification and validation of new tumour markers [55–57]. Among them, individual Se-species could potentially reflect the complex relationship between Se status and functional alterations occurring in cancer cells.

About 25 Se-proteins have been identified in the human proteome [1]. The most interesting as biomarkers are blood/serum Se-proteins, that could accurately reflect the global Se status at short term and are also easy accessible for sample collection. Serum GPx3 is the only extracellular enzyme of a family (GPxs) comprising at least 6 isoenzymes playing general antioxidant functions. They protect the organism from oxidative stress by catalyzing the reduction of reactive oxygen species (ROS) at the expenses of glutathione. GPx3 has been proposed to be a major scavenger of ROS in the extracellular space [58]. Serum SeIP is a secreted glycoprotein that has been hypothesized to be responsible for the transport of Se through the whole body. It is the only Se-protein containing ten residues of SeCys (instead of one) and has few hours (3–4 in rat) of half-life in plasma [59]. The third species of Se in serum is the SeMet-containing protein SeAlb. Even if SeAlb could have no direct biological functions, the unspecific incorporation of SeMet into body proteins like albumin allows Se to be stored and reversibly released by normal metabolic processes [60].

Considering their properties, serum SeAlb and consequentially GPx3 concentrations were expected to decrease as an effect of the high oxidative stress level typical of cancer. However, in our study no significant differences between Se-proteins level in CRC patients and the control group were found. It is worth noting that CRC patients constitute a heterogenic category, characterized by wide variability of Se status during the progression of disease. When CRC patients were split according to the cancer stage, SeAlb level resulted significantly higher in TNM I patients than in the TNM IV or H groups, while GPx3 level was decreased. This result suggests a possible association between SeAlb (and GPx3) and CRC progression. It is difficult to explain the enhanced SeAlb level in serum of TNM stage I patients. It is not clear yet if this behaviour is a cause or a consequence of CRC, nor what are its biochemical bases, but it might be interpreted as an anomalous Se-proteins system regulation. In this context, it is possible to explain the subsequent re-normalization of all Se-proteins level in the more advanced cancer stage (TNM IV) as the effect of a further increase of Se-proteins system stress, which produces the annulment of the response to cancer. Such explanation is compatible with the general model of the anticarcinogenic action of Se proposed by Combs et al. [61]. On the other hand, since SeIP plays mainly the role of a carrier of Se through the whole body its concentrations in response to increased oxidative stress is a complex function of the balance among depletion rate, production rate and the distribution velocity. This could explain why SeIP level solely is not significantly correlated with CRC in our study. However, the relatively higher concentration of SeIP in TNM stage I patients confirms the importance of this Se-protein in the hierarchy for Se incorporation, in respect to SeAlb and GPx3.

SeAlb level tended to reduce with the age in healthy persons, an effect that was not observed for GPx3 and SeIP. Other studies noticed the same behaviour for SeAlb or for total Se [8,62]. This association has been ascribed to a less efficient absorption or increased elimination of Se with aging [63]. We did not observe the same effect for CRC patients, supporting the hypothesis that presence of CRC alters the homeostatic capability of the Se-proteins system.

It is important to state that another potential confounding factor for the Se-proteins distribution (particularly SeAlb) is Se dietary intake, that was not considered in this study. However, the correlation between individual serum Se-proteins and dietary intake is very difficult to assess and has still not been clearly demonstrated [64–66]. Since all the subjects recruited in this study were resident in the same non deficient Se region (80% in Veneto, Italy and 20% in adjacent areas) [67] and underwent the same intestinal preparation during 2–3 days before blood collection, we assumed that such a confounding effect is negligible.

SeIP and SeAlb were interestingly associated with some prognostic criteria, which were investigated here for the first time in relation to Se. Peritumoral lymphocytic infiltration is a histopathological marker given by a spontaneous influx of B and T-lymphocytes in the tissue surrounding the neoplastic region, probably promoted by the secretion of antigens or other factors by the neoplastic cells. Several evidences have proven that the presence of lymphocytes around CRC tissue correlates with a less aggressive behaviour of the tumour, and thus is considered a significant positive factor in prognosis effectiveness and survival of CRC patients [68]. The decrease of SeIP and SeAlb levels we observed in patients with infiltration is compatible with the effect of Se in improving the activation and proliferation of B-lymphocytes and enhancing the T-cells function [69]. This result further supports our hypothesis that an increase of SeAlb level in CRC patients is representative of a suppression of the host’s immune reaction to the tumor. More difficult is to interpret the correlation we observed between SeAlb and tumour grading. This parameter express the level of morphological differentiation and mitotic rate of neoplastic cells in respect to the normal tissue. Grading is an index of cancerous cells aggressiveness, and thus is negatively correlated with good prognosis [70]. The parameter is influenced by a complex combination of characteristics of both neoplastic cells and metabolicism of the host (in turn dependent by the age). Thus, its association with SeAlb could be more indirectly related to unknown metabolic pathways. Even more challenging is to forward an interpretation for the slightly higher level of SeIP when LOH was evidenced. Loss of heterozygosity implies the presence of micro deletions on specific regions of the 18q chromosome, where candidate tumour suppressor genes are located. Such phenomenon is believed to be one of the key steps to carcinogenesis of CRC, and is also considered a potential prognostic factor. On the bases of our preliminary results, deeper investigations could be carried out to elucidate the specific biochemical connection between serum Se-proteins level and the immune/ metabolic mechanisms in CRC.

Data mining technique performances showed the interesting potentiality of the complete serum Se-proteins pattern as a combined
biodmarkr for early cancer stage detection, even if the models need to be validated over a larger sample size. According to most of the previous studies on these techniques, the method with the best classification performances differs for each data set and structure [71]. This aspect is critical in our pilot study, where the low sample size and variables space stress the models to be very simple. Artificial neural networks methods, in particular PRAANNs, emerged as the most powerful for pair-groups classification. The main advantage of these techniques is that they are not forced on a linear structure, and thus can model a wider range of data structures [48]. On the other hand, ANNs operate like an empirical “black box” which topology could have no relation to the real mechanisms of the modelled system. LRA techniques are based on linear functions and were effective only when the groups to be discriminated differ markedly, as was the case of TNM I vs. H comparison. Classification trees were too simple to perform satisfactorily in our study. The good performance of ANN models demonstrates that serum Se-proteins can be accurate predictors for early stage CRC. This confirms but also strengthen those associations were ascertained from the t-tests for pair-groups comparison. Still, by comparing the performance of LRA, CTs and ANNs it is possible to infer that the serum Se-proteins system interact with CRC in a relatively simple but nonlinear fashion.

Data mining in general allows to identify unknown patterns, relations and meanings that could not be obtained by applying traditional statistical methodologies alone. Each classification technique presents peculiar strengths and weaknesses in respect to the data structure and size, noise presence, missing data and results interpretation [72]. Our results point out that in Se-proteins/biomarkers investigation, a set of data analysis techniques should be employed instead of a single method in order to explore different data structures, particularly when complex and unknown biochemical mechanisms are involved.

5. Conclusions

This study had revealed that early stage CRC could be associated to a variation of serum Se-proteins concentration, in particular SeAlb. Promising correlations were also detected between Se-proteins and clinical parameters including peritumoural lymphocytic infiltration and grading, that are important prognostic factors for CRC. The results mark the importance to move from total Se to individual Se-species association among Se-proteins status in serum and tissues, presence of cancer and prognostic criteria as well as determination of proteins containing essential trace elements other than Se. The best biomarker for CRC is expected to be not an individual (Se-protein) variable but rather a selected group of proteins. Work is under progress to select, for CRC is expected to be not an individual (Se-protein) variable but rather a selected group of proteins. Further studies should be also carried out in order to investigate the causal relationship between CRC and serum Se-species, which could provide more accurate epidemiological and nutritional data in order to enhance the quality of prevention and treatment for CRC. The systematic introduction of speculation analysis and multidisciplinary approach involving analytical chemistry, oncology and statistics that we propose emerges as a powerful strategy to pursue these challenging issues.

Acknowledgments

This study is a contribution to the Marie-Curie Intra-European Project (MEIF-CT-2006-024156/ELSA-BIM) funded by the European Commission. ELGA LabWater is acknowledged for providing the PURELAB Option-R and Ultra Analytic systems, which produced the ultra-pure water used in these experiments.

References
