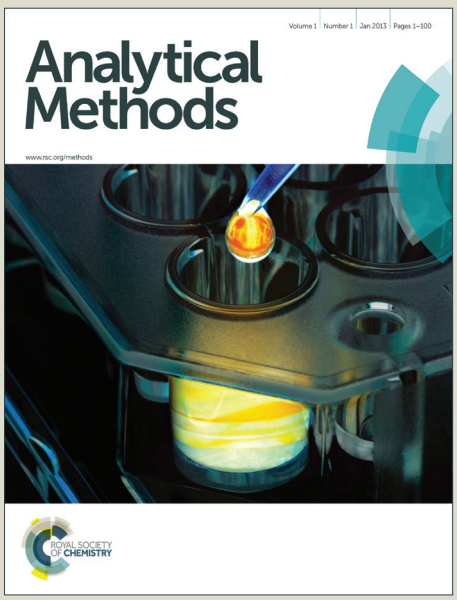
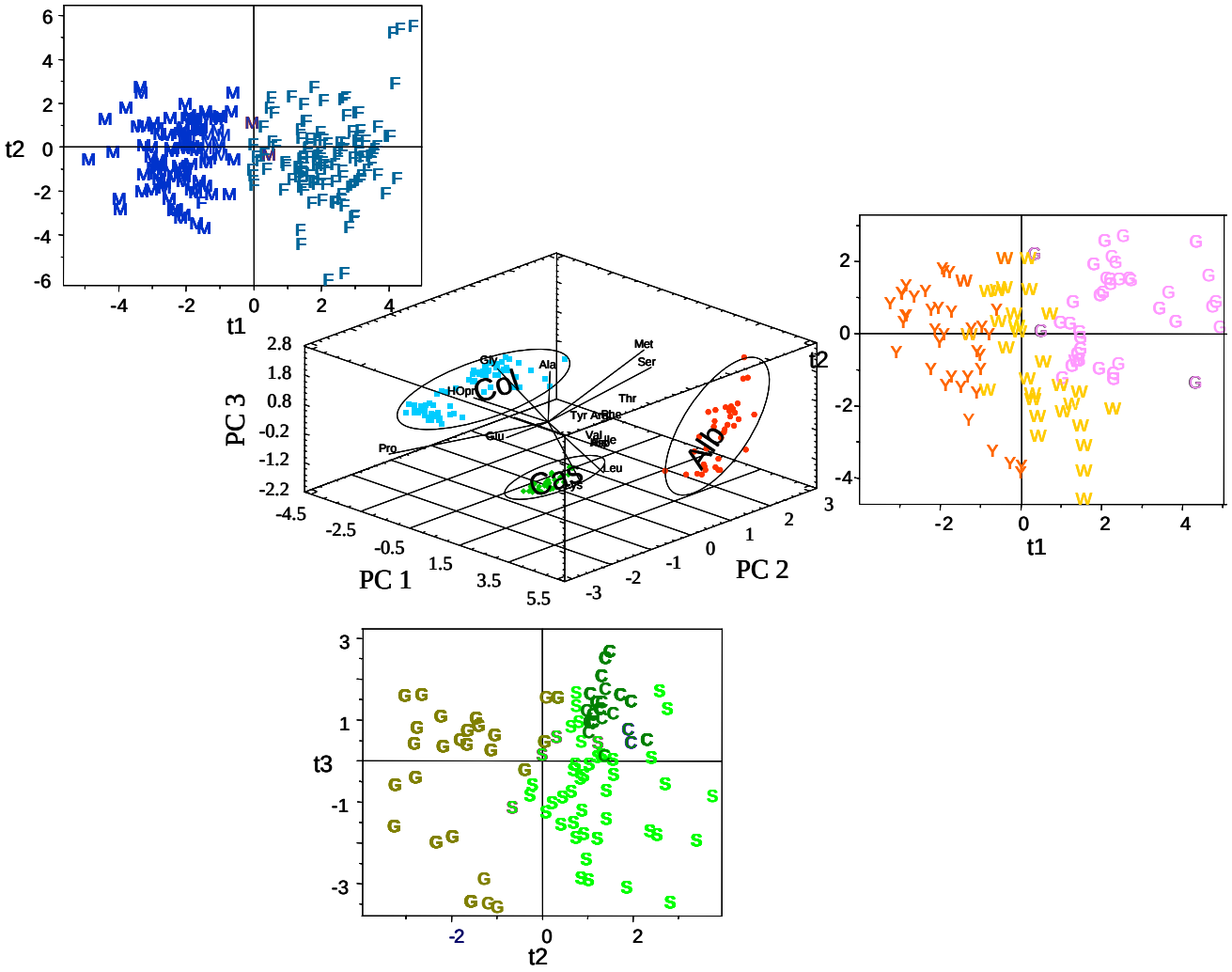
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods



CHARACTERISATION AND CLASSIFICATION OF BINDERS USED IN ART MATERIALS AT CLASS AND SUBCLASS LEVEL

R. Checa-Moreno[1], E. Manzano[2], L.F. Capitán-Vallvey[2 *]

[1] *Laboratorio Central de Sanidad Animal, Ministerio de Medio Ambiente y Medio Rural y Marino, Camino del Jau s/n, E-18320 Santa Fe, Granada, Spain.*
[2] *ECsens. Department of Analytical Chemistry, Campus Fuentenueva, Faculty of Sciences, University of Granada, E-18071 Granada, Spain.*

Abstract

SIMCA pattern recognition is used with amino acid chromatographic profiles in a large homemade collection of natural protein binders obtained following old recipes traditionally used by painters and considered here as the standard of classification. An initial cluster analysis of the full data set made it possible to distinguish three main classes of protein binders: albumin, casein and collagen-like substances. An additional iterative study of each class revealed a new subclass, i.e., glair, yolk and whole egg for the albumin class; goat, sheep and cow for the casein class; and mammals and fish for the collagen class. Optimized SIMCA models for each class and subclass were obtained with good results in terms of sensitivity (90-100 %), specificity (73-100 %) and interclass distance (>1.4), providing identification of the protein binder present in a set of samples of different origins such as natural products, commercial binders and works of art considered cultural heritage.

Keywords: Amino acids, Liquid Chromatography, Soft Independent Modelling of Class Analogy Pattern Recognition, Protein binder, Two-level classification.

1. Introduction

The organic binders used by artists in the preparation of a painting determine the artist's technique, and differentiate painting styles [1]. Moreover, the knowledge of the kind of binder present can help specialists to authenticate or refute questionable works of art. Artists over time have used a wide variety of procedures preserved in recipes to improve and/or modify the painting properties of materials. The origin of the binding media present in the pictorial layer of artworks is a question in the analysis of cultural heritage materials that has not been resolved. This information is necessary to establish the historical provenance of materials from among schools of art and even to authenticate or refute questionable works of art.The substances used include drying oils, resins as components of varnishes, sugars, proteinaceous materials and waxes, among many others, and also complex types of mixtures of them. Since ancient times, the proteinaceous materials used as binders in the colour layers of old paintings have been found in nature and include: animal glues prepared from animal skin or bones containing several types of collagen, egg white, egg yolk and casein [2].

The identification of both the chemistry and origin of proteinaceous binders is not an easy task for several reasons as reported in the literature: a) they are natural products and artists obtain them using old recipes usually without any prior purification steps; b) the proteinaceous materials found in paintings are used either alone, in combination with oils or with other organic materials such as impurities resulting from their preparation; c) the organic materials tend to suffer degradation, chemical transformations and oxidation processes with the environment, pigments and others substances that can change their initial chemical composition by aging and degradation processes; [3-5] d) the small amount of sample available and occasionally the small percentage of binder. In addition, the difficulty in identifying them is exacerbated by the fact that the artists might have used mixtures of

several types of organic materials and sometimes undocumented formulations in their search for artistic effects and mechanical behaviours which they use to give shape to their work [2].

The great variety of analytical methods proposed in the literature, sample treatment procedures, strategies and mathematical tools for data treatment have made it possible to discriminate among oils, proteins and other classes of binders [6] and, although with more difficulty, between the three types of proteinaceous materials used as paint media, i.e. egg, casein and collagen [1,7,8]. The earliest works that identified protein binders were based on the use of observational methods with stratigraphic cuts of pictorial samples based on coloured or fluorescent reactions [9], solubility tests [10], immunological techniques [11,12], and more recently immunodetection-based methods [13], although these have not yet been adapted to routine analysis in conservation laboratories. The classic analytical methods make it possible to discriminate between the general categories of binding media (oil, gum, protein, wax and terpenic resin) by qualitative means. Different optical instrumental techniques such as FT-IR [4,14], diffuse reflection infrared spectroscopy [4], Raman and micro-Raman spectroscopy [4,5,15-18], and NMR[19] have proven useful in the study of artworks because of their versatility in obtaining analytical information from both inorganic and organic materials and also performing ageing studies. Nevertheless, so far the characterization of organic binders, in particular proteinaceous materials, has been essentially performed using chromatographic techniques [20-22]. The first chromatographic techniques, both paper (PC) and thin layer (TLC), have been progressively replaced by high performance liquid chromatography with fluorescence or UV-Vis detection[20], gas chromatography (most commonly used with mass spectrometry detection (GC-MS),[1,23,24] coupling analytical pyrolysis (Py-GC-MS) or a wet-chemical treatment of the samples prior

to CG-MS analysis (chemolysis and derivatization reactions) and capillary electrophoresis[25]. Recently, proteomic techniques mainly used for biological sample analysis have been introduced for painting analysis [7,26-28] once they were adapted to handle the requirements presented by the specific situation. Proteomic approaches based on mass spectrometry applied in conservation science have promising results for identification of the binder protein in mixtures mainly at a group level, i.e. with egg, animal glues and milk products. Only limited results in conservation science have recently been published: the distinction between egg yolk and egg glair temperas [29], different milk species [30], and animal glues [31] have been studied to some extent. This method also solves the outstanding problem of the identification of the mixtures of proteinaceous binders, which is typical for the other commonly used analytical methods but not that of identifying/discriminating the source of proteinaceous binders.

In the field of cultural heritage, the identification of the categories of proteinaceous materials through their amino acid composition is based on the evaluation of the some chromatographic amino acid profiles or the presence of specific markers, making it possible to differentiate between eggs, casein and collagen used as paint media. Over time, the strategies have increased in the number of amino acids used to make the identification and consequently the complexity of data treatment [6] and has become more robust. Several strategies have been developed: (a) amino acid ratio flow charts [21]; (b) bidimensional plots of amino acid ratios [32]; (c) joint amino acid profiles of the sample using a correlation index estimated with amino acid profiles of samples and standard databases [33], (d) multivariate statistical analysis such as principal components analysis (PCA) [34,35], factor analysis (FA)[1] and neural networks [36], (e) use of multivariate approaches based on the SIMCA technique [8].

All the strategies considered have in common the fact that they use reference proteinaceous standards. This is another important aspect be taken into account. The correct identification of protein binders by comparison with reference proteinaceous standards will depend on selecting the standard used well. Many researchers have used chemical standards of purified proteins to perform identification but, as mentioned above, the protein binder material present in an artwork sample is an entirely natural product and consequently a very complex substance, so it is important to use standards of natural products similar to those used by artists in the past. Additionally, the intra-specie variability must be taken into account by considering the protein binder standards from different individuals belonging to the same or different species.

This paper presents some significant results obtained from the use of the soft independent modelling of the class analogy classification technique (SIMCA) [37] on the profile of amino acids collected by HPLC-DAD analysis. Both reference materials and samples from works of art have been analyzed using phenylisothiocyanate (PITC) as the derivatization reagent [20]. Amino acid profiles were obtained from a collection of reference proteinaceous binders prepared by us and a test set from paintings, manuscripts and sculptures from the 15–18th centuries. With SIMCA, more than with traditional strategies, it is possible to use software to know the confidence level for each classification made. This is performed by an appropriate statistical F-test. The strategy is important to differentiate the painting technique adopted by different artists and is useful for classification purposes and provenance studies.

2. Materials and Methods

2.1. Reagents and solutions.

All chemicals were of analytical grade. Individual standard amino acids analyzed were purchased from Sigma (Deisenhofen, Germany), phenylisothiocyanate (PITC) and triethylamine (TEA), hydrochloric acid, acetonitrile (HPLC quality) and acetic acid were obtained from Panreac (Montcada i Reixac, Barcelona, Spain) and absolute ethanol from Merck (Darmstadt, Germany). Standard stock solutions of each amino acid were prepared by adequate weighing and disolution in 0.1 M hydrochloride acid (HCl). Reverse osmosis quality water was produced by a Milli-RO and Milli-Q 185 Plus purification system (Millipore Co., Bedford, MA, USA).

2.2. Standards of natural protein binders.

A collection of 143 natural proteinaceous binders traditionally employed by past artists was prepared (Table 1) and used as classification standards[38]. Egg protein standards were prepared from whole eggs or by physically separating the glair and yolk. Standards of casein were prepared from previously skimmed milks by centrifugation at 30000 R.F.C. and subsequently acid precipitation to pH 4 with hydrochloric acid, at room temperature. Collagen standards were obtained of fish skins, backbones and air bladders and mammal skins, bones and cartilage from different species by lixiviation in boiling water. Approximately 2 mL of each protein binder natural standard were individually aliquoted in 5 mL vials, dry-frozen and conserved by freezing at -20ºC for the correct long conservation.

Table 1

2.3. Apparatus and software.

A Pico Tag workstation from Waters (Milford, MA, USA) for protein hydrolysis and amino acid derivatization provided with an oven (100-150º C) was used. A Hewlett-Packard HP

1090 liquid chromatograph (Palo Alto, CA, USA) provided with a Diode Array Detector (DAD) and an Aminoquant ODS column (5 µm 200 x 2.1 mm i. d.) was used. The PITC derivatives were identified by their retention time at 254 nm. The chromatographic conditions for amino acid determination as PITC-derivatives were those previously optimized by us [20]; column heater: 40°C; flow-rate 0.5 ml/min; buffer A: 0.28 M sodium acetate, 0.075 % (v/v), TEA, and 6% acetonitrile (pH 6.38); buffer B: 60% acetonitrile. Mobile phase gradient was: 0% B at 0 min, first linear gradient 2% B at 2 min, second linear gradient 43% B at 9 min; 50 % B at 13 min.

The Mettler AE 160 analytical balance used (Mettler-Toledo AG, Greifensee, Switzerland) was regularly checked with certified type E2 weights (5 mg, 100 mg and 100 g). The fixed volume micropipettes (Biohit, Helsinki, Finland) were periodically controlled through gravimetry to ensure the traceability of the results.

For treatment and later data analysis, the software packages Statgraphics Plus for Windows by Statistical Graphics Corp. and SIMCA-S for Windows ver. 5.1 (1994) by Umetri AB (Umea, Sweden) were used in a Pentium 300MHz personal computer. The SIMCA-S software package included modules to define a data file; to scale, weigh and transform data; to edit and list the files; to input the data, define classes and perform principal component analysis for classes; to test the fit of data to defined classes; to perform several plots as PC-scores, loadings, etc.

2.4. Analytical procedure.

A small amount of standard protein binder or test sample (1-10 mg) was dissolved in 0.05 M pH 12.3 phosphate buffer solution and 25 µl of this solution subjected to hydrolysis and PITC derivatization according to the Waters Picotag© method. Before sealing the samples

in a vacuum for hydrolysis at 110°C for 16 h, the dry samples in small tubes (6 x 50 mm) were placed in the reaction vial with 200 µl of 6 M HCl. The hydrolyzed samples were dried and redried by adding 20 µ1 of ethanolic solution (ethanol-water-TEA) to ensure that a trace amount of ammonia was left. For derivatization, the samples were coupled with 20 µ1 of PITC solution (ethanol-water-TEA-PITC, 7:1:2:1) for 10 min, dried again in the workstation, and reconstituted for analysis in sample diluent (0.5 M sodium phosphate buffer, pH 7.4, and 5% acetonitrile). The total amount of each amino acid for each standard or test sample was determined (in picomoles) by a weighted calibration based on the peak area to internal standard ratio. Each protein binder standard was analyzed by three-five replicates in conditions of reproducibility in order to consider the variability of the sample treatment method.

3. Rationale

SIMCA (Soft Independent Modelling of Class Analogy) [39] is a supervised classification technique that builds a distinct confidence region around each class. A principal component analysis (PCA) is performed on each separated class in the data set, and a sufficient number of principal components are retained to account for most of the variation within each class. New objects are considered to belong to the class if their Euclidean distance towards the constructed PC space is not significantly larger than the Euclidean distance of the class objects towards their PC space. The variance that is explained by the class model is called the modelling variance, which describes the signal, whereas the noise in the data is described by the residual variance or the variance not accounted for by the model. By comparing the residual variance of an unknown $S_x^2(q)$ to the average residual variance of

those samples that make up the class $S_o^2(q)$ by an F-test it is possible to obtain a direct measure of the similarity of the unknown sample to the class.

$$F = \frac{S_x^2(q)}{S_o^2(q)} \qquad \text{Eq. 1}$$

An advantage of SIMCA is that an unknown is only assigned to the class for which it has a high probability. If the residual variance of a sample exceeds the upper limit for every modelled class in the data set, the sample would not be assigned to any of the classes because it is either an outlier or comes from a class that is not represented in the data set. There are diagnostics to assess the quality of the data, such as the modelling power (MP) and the discriminatory power (DP). The modelling power describes how well a variable helps the principal components to model variation, and discriminatory power describes how well the variable helps the principal components to classify the samples in the data set. Variables with low modelling and discriminatory power are usually deleted from the data because they only contribute noise to the principal component models and new models with lower variables are developed again.

When several classes are present, it is of interest to have a measure of the distance between each pair of classes. This can be calculated as, for instance, the pooled variance of the residuals obtained when objects of class "one" are fitted to the class model "two", divided by vice versa the pooled residual variance obtained when the objects are fitted to their "own" class model. Suppose two class q and r, are to be studied. Two different models will be constructed for each class. The distance between two classes, $d_{(r-q)}$ is calculated by eq. 2, where $S_r^2(q)$ is the residual variance of class r fitted to class q and $S_o^2(q)$ is the variance within class q. When the distance between two classes is close to zero, the classes are very similar; values near to 1 indicate poor separation and values larger than 2 good

resolution. This distance can be compared with F-statistics to judge the significance of the class separation [37].

$$d_{rq}^2 = \frac{\sum_{j=1}^{p} \left[s_r^2(q) + s_q^2(r) \right]}{\sum_{j=1}^{p} \left[s_o^2(q) + s_o^2(r) \right]} \qquad \text{Eq. 2}$$

In this work, the class models have been developed with a higher number of objects and using the interclass separation as criterium of optimization, and consequently with a different combination of variables/objects to those employed in our previously published paper[38]. Data analysis is performed in a few steps: a) preliminary univariate data analysis to detect possible outliers, information about the relevance of variables, etc.; b) cluster or principal component analysis of the complete data set to establish classes, groups, clusters, etc.; c) SIMCA model development of the emerging groups; c) Optimization of SIMCA models by deleting outlier objects and noise variables. This can be achieved by choosing variables which contain the largest amount of modelling or discriminant information for the classification. After deleting irrelevant variables or outliers, the new PC models are refitted.

4. Results and discussion

4.1. Homemade protein binder collection.

The starting condition to build a model to classify protein binders by origin is an arrangement of a set of samples with enough specimens. To cover a wide variety of traditionally used protein binders, several albumin, casein and collagen-like species were considered to build a collection of reference substances. At least two specimens belonging to the same species was obtained whenever possible in order to consider the intra-specie variability. The standard preparation for the proteins was done using old recipes which

produced the standard substances similar to those used by past artists. Current knowledge about the techniques used over the centuries in the creation of artworks comes mainly from historical treatises that provide an overall view of the techniques used in different places and ages. The book *Il Libro dell'Arte* by Cennino Cennini [40], written at the beginning of the 15th century and considered a practical handbook describing common techniques from the late 13th and mid-14th centuries, was used as reference for preparing the samples.

Eggs and natural milks were obtained from local farms while collagen-like substances were collected from different parts of fish and mammals that had been previously purchased in different slaughterhouses and supermarkets in Granada (Spain). At least two eggs belonging to each of the species considered were collected and used as egg protein standards, one egg to prepare the whole egg standards and the other to obtain the glair and yolk standards. Skimmed milk samples from different species and origins were acidified to precipitate the casein fraction. This more artificial way was preferred to the classic way of naturally skimming milk by letting it settle followed by lactic-induced precipitation of the casein fraction because it is avoided some microbiological contamination and later degradation. Human and donkey milk samples were collected from lactating mothers to increase the variability of casein group. Human samples were provided by fully lactating, healthy mothers during the first stage (1-5 days) of lactation. Finally, eighty-one collagen-like standard samples were obtained of skins, bones and cartilages from mammals and skins, backbones and air bladders from several species of fish, by a lixiviation process in water.

4.2. Data analysis.

Figure 1 shows the chromatogram types of the albumin, casein and collagen samples. There are several clear differences among the three kinds of substances. HOpr is an amino acid only present in collagen-like substances and is therefore useful when differentiating between collagen and albumin or casein substances. The contents of Asp, Glu, Ser, Phe, etc. are also interesting in terms of discriminating albumin from casein substances. The problem is more complex when distinguishing between several substances containing the same principal protein, for example glair, yolk or whole egg substances, since all of them belong to the albumin-like complex where the amino acid profiles are very similar. Therefore, using a few chromatographic peaks for differentiation may not provide enough confidence due to the similarity of these proteins in structure and properties. The application of multivariate statistical methods is thus helpful since it works with the overall amino acids (peaks) and their rates, establishing the differences.

Figure 1

All standards/specimens from the natural collection of proteinaceous binders were analyzed using 3-5 replicates. In this way, data obtained contained the variability in each natural species resulting from its genotype differences and also to observe any error in the analytical method. The amino acid composition of samples has been discussed in the vast literature in several forms including column mass injected, molar and mass percentage, etc. Here the raw data obtained were described in a pMol-injected on column basis but the data generated were subject to a process of internal normalization consisting of the expression of the contents of each individual sample as a percentage of the sum of its amino acids. (The full data for all the samples are available as Electronic Supplementary Information, ESI Table S1). This process is appropriate for many characterization problems in which the shape of the profile signal, and not the intensity, contains the relevant information.

However, the quantitative information is lost. We selected the molar percentage, although the absolute amino acid content in several replicates samples may vary depending on the error in weighting the sample and solution volumes used in the sample treatment.

Univariate analysis. Univariate analysis was performed on the full raw data set (455 rows x 18 columns). The Box-and-Whiskers plot analysis highlighted one outlier in the collagen-like class (No. 112, ESI Table S1). Since none of the samples showed outliers for more than three-four variables of the seventeen used, neither sample was rejected *a priori*. To establish the discriminant capacity of each amino acid, a one-way ANOVA using Fisher's least significant difference criterion (LSD) at 95% confidence level was performed using the species as the criterion to compare the mean values (ESI Table S2). It concluded that there were many amino acids, making it possible to completely differentiate the three main classes: albumin, casein and collagen. Only HOpr, His, Leu and Lys showed no statistical difference in distinguishing between albumin and casein classes and, analogously Arg and Pro in distinguishing between albumin-casein and casein-collagen, respectively. The case of HOpr for differentiation between albumin and casein classes is obvious because this amino acid is not present in these kinds of proteins, but the HOpr composition in these protein standards was written as 0.2 pMol, i.e., the detection limit. There were several amino acids that completely differentiated the three kind of proteins considered. Additionally, the ANOVA analysis was performed to check for the possibility of differentiation between subclasses according to their origin but good results were not obtained and consequently we resort to pattern recognition techniques.

4.3. Principal component analysis.

PCA makes it possible to visualize data set information in a few principal components, retaining the maximum possible variability. Scores for each sample and loadings are represented in the three first principal components in Figure 2. In the first principal component, the collagen samples to the left of the graph that have a negative score are completely separated from the remaining samples. To the right of the graph, albumins and caseins can be separated along the second component. The casein samples have higher scores than the albumins. The albumin class shows higher score dispersion than the remaining two classes. Similar results were obtained from the clustering analysis.

Figure 2

Other important information obtained from the principal component analysis is the loading plot. The variables responsible for the separation of two classes can be directly identified. An examination of the variable loadings from the principal component analysis showed that the contents of Gly, HOpr, Glu, Ala and Pro were the most responsible for the formation of the collagen class, whereas the greater contents of Ser, Thr and Met were for albumin class. Finally, the casein class can be differentiated from the albumin by the content of Lys.

4.4. SIMCA class modelling.

SIMCA is a modelling technique that builds a model for each category or class. The centre of the model is the mean value of the objects and the space orientation is defined by the principal components. A range for each component is built on the basis of the score distribution. A scale effect in raw data can be avoided by scaling the variables. The most common way of doing this is using the z-transform, also called autoscaling. This refers to mean-centring followed by dividing by the standard deviation for each sample. This

produces a feature with zero mean and a unit variance. Multivariate analysis was performed on the autoscaled data. The good separation of the three classes observed in the PCA plot made it possible to construct the SIMCA models. The first objective was to find PC models that would separate the three kinds of protein substances. The SIMCA analysis with all the variables showed that the classes (albumin, casein and collagen) can be well described by PC models with two, three and three components, respectively. The explained variance for each model is 56, 67 and 64 % and the sensitivities are 90, 85 and 89 %, respectively with an excellent specificity of 100 %. On the basis of low modelling power (MP) and low discriminatory power (DP), several variables and objects showing a high leverage effect were deleted from each class. The new models obtained on the basis of the remaining variables and objects are described in Table 2.

Table 2

Nine, thirteen and eleven objects were considered as outliers of the albumin, casein and collagen classes, respectively. HOpr, an amino acid not found in the casein protein, was used due to the z-score scale transformation of the data employed. The three class models showed very good sensitivities and full specificity. SIMCA also provides differentiated information about the variables through the modelling and discriminant power. The modelling power is the contribution of each variable to the model and the discriminant power is the capacity to differentiate among classes. All amino acids have a similar modelling power around 0.5 in the albumin class, as do Glu, Pro, Val and Leu in the case of casein, and the most hydrophilic as acids, Asp, Glu, Ser and Gly, are the highest modelling variables in the collagen class. Regarding discriminant power, Glu and Gly are the most discriminant amino acids between the casein and collagen classes; HOpr, Gly and Leu

between the albumin and collagen, and Glu and Arg between the albumin and casein. These three class models were perfectly separated.

4.4.1. Subclass analysis.

Figure 2 shows that the three categories of binders are well separated in the principal component space, but it is not clear if any distinction can be made according their origin in each category. To investigate this, and taking in account the great number of objects available for each category, a new separate one-to-one PC analysis was performed for each class. Figure 3 presents the results obtained. Three new subclasses can be distinguished in the albumin class according to the egg fraction prepared: glair, yolk and whole egg; for casein according to the taxonomical family Bovidae: bovinae, caprinae and genus ovis (goat, cow and sheep); and for collagen two new subclasses related to the class of Subphylum Vertebrata (mammals or fish). PCA can find new subgroups when a high number of objects are available. Obviously, this ability is not due to the data treatment systems, i.e, this is not something inherent in PC analysis, but is due to the proper nature of the problem. New classes, namely subclasses, can appear because the objects belonging to the subclasses have singular properties. These new properties make it possible to differentiate among subclasses. The real virtue of the methodology (PCA, SIMCA, etc.) is finding the new subclasses on the basis of the data set available.

Figure 3

4.4.2. Subclassification of the albumin class.

Figure 3a shows that new subclasses are well differentiated in first principal component. Glair objects have positive scores whereas the yolk subclass is negative; obviously the whole egg, as a mixture of glair and yolk, lies between them. New PC models for these subclasses were built by deleting noisy aminoacids and following an iterative optimization

process based on specificity, sensitivity and especially inter-class distance criteria in order to guarantee a good separation. The submodels are summarized in Table 2. The amino acids used to model the yolk subclass were the most hydrophobic which fits with the fact that the lipoproteins such as phosvitin, livetin, lipovitellin, etc. present in yolk egg are made up of lipophilic amino acids. The amino acids HOpr, His and Met were not used to generate either of the models of the subclass because their low modelling and discriminating power. This was in agreement with the fact that the cluster analysis previously performed on the variables applying Ward's clustering method, with the Euclidean distance as the similarity measure, presented two groups (Figure 4): the first brings together the principal amino acids used in the modelling and the second contains the amino acids HOpr, His, Met and Tyr with no participation in any of these submodels. Note that HOpr is not present in these proteins and His, Met and Tyr are the most irreproducible amino acids in the hydrolysis step in the used Pico-Tag method.

Figure 4

A measure of the distance between two classes r and q is calculated from a) the total residuals obtained when all objects in class r are fitted to class model q and vice versa all objects in class q are fitted to class model r in comparison with b) the residuals when all objects in class q and r are fitted to their "own" class models. Table 3 gives the class distances for (i) the innitial models with all the variables and (ii) the optimized models with the retained variables. In both cases the subclasses are fairly well separated (d>1) and the separation increases when the irrelevant variables are deleted. The whole egg subclass is also visibly closer to glair than yolk. Acceptable distances were obtained between the whole egg and both glair and yolk (2.5 and 3.4, respectively). The distance between the glair and yolk was the highest (6.4) as well.

Table 3

In order to find another criterion of classification inside albumin class news PC analysis were performed. The PC projection of the egg glair samples codified according their phylogenetic origin made it possible to distinguish them in the space of the first principal component. Glair samples are phylogenetically separated at the level of order or family. All birds belong to the Animalia Kingdom, Phylum of Chordata, and Class Aves (birds). At the order level, the birds begin to diverge: Anseriformes (ducks, geese, screamers, swans, and waterfowl), Coliiformes (mouse birds and colies), Columbiformes (pigeons and doves), Galliformes (chickens, fowl), Pisciformes (woodpeckers) and so on up to at least twenty-three orders. Glair sample projection shows the two well-defined groups: samples belong to the orders Columbiformes and Galliformes (G&G) and, on the other hand Anseriformes samples (A). Table 4 shows the features of the new SIMCA models developed. Threonine, aspartic acid, serine and glutamic acid are the amino acids with the greatest discriminant power between the A and G&C classes. In other words, the amino acids with high polarity are responsible for distinguishing between the two classes considered here. The statistical interclass distance was 4.2 (>1), therefore showing a good separation between the two classes. Figure 5 shows the Coomans plot of the two SIMCA models. None of the models built admitted samples from the other class and the specificity was 100 %.

Table 4

Figure 5

The same approach with egg glair was carried out with the yolk and whole egg samples. It was not possible to find a similar behaviour as with the egg glair. A good separation for yolk and whole egg according to the phylogenetic origin was not found, perhaps because the amino acid composition of these egg fractions is influenced by the great number of

protein substances present in egg yolk. These protein substances may introduce a hidden effect in the amino acid profile of the yolk and whole egg samples.

4.4.3. Subclassification of the casein class.

As with the albumin class, the PC analysis was performed with the objects from the casein class, revealing the appearance of three new subgroups or subclasses: sheep, goats and cows. Figure 3b shows that the new groups are well differentiated. The sheep objects have positive scores in the second PC whereas the goat objects are negative; cows with positive scores are separated from the sheep along the third PC. The new submodels optimized are shown in Table 2. It can see that amino acids with high polarity such as Asp and Glu were not used to model the cow subclass. The goat and cow models showed a full specificity whereas the sheep model reported 81%; this was because six and two objects that belonged to the goat and cow classes respectively were incorrectly assigned to the sheep class. The best separation between the sheep and cow classes was obtained when the most polar amino acid (Asp to Thr) was present only in one of them. For that reason, the polar amino acids Asp and Glu were used in the sheep class but not in the cow class. In the same way, HOpr, Ser, Gly and His were only used to model the cow subclass but not the sheep subclass. The modelling power of the variables retained was very similar. None of the amino acids proved to be especially significant in modelling these subclasses. With respect to the discriminant power, the most significant amino acids were HOpr, Ser, His, and Tyr. Separation between the goat-sheep subclasses is due to Tyr, Gly and Met. Ser, His and Tyr were the most important amino acids in the goat-cow differentiation. Finally HOpr, Tyr and His had a role in distinguishing between sheep and cow. Tyr is a very important amino acid in the separation of the three casein subclasses.

4.4.4. Subclassification of the collagen class.

As with the albumin and casein classes, PC analysis was performed on all the objects belonging to the collagen class in orther to find new subclasses. Even though the collagen standards were obtained from several animal parts, such as skins, backbones and air bladders for fish and skins, bones and cartilages for mammals, the new subgroups appeared when the samples were codified according to their phylum membership: mammalian and fish. No separation was observed when either fish or mammal samples were projected individually on PC plots (in other words it was not possible to distinguish between fish and mammal samples using the animal part as criteria). Figure 3 shows that the new subclasses are well separated. Mammal objects have negative scores on the first PC whereas fish objects are positive. The new PC models for these subclasses were optimized as shown in Table 2. The non-polished models (A=3) with all the variables and available objects for mammalian and fish subclasses reported: 60 and 57 % of variance explained, 86 and 88 % as sensitivity and 100 and 90 % specificity, respectively; with a good interclass distance (2.1). The optimization performed for the sub-models produced tabulated results. His, Pro, Val and the lowest polar amino acids Leu, Phe and Lys did not participate in the modelling. The fish model was built using the most polar amino acids (Asp-Gly), Ala and Met. On the other hand, Ser, Met and Thr were the more important amino acids and made it possible to discriminate between mammal and fish subclasses. This fits the data published by Eastoe.[41] Not enough information is obtained from the amino acid profile to differentiate the samples inside the mammal subgroup according their species (bovine, porcine, ovine) [42] or in the fish subgroup according species or animal part (skin, bone, air bladder).

This methodological strategy supported by the development of robust class models generated using natural standards similar to those used by the old masters presented here

makes it possible to identify the hierarchical nature of protein binders at different levels (class and subclass) in a non-subjective way.

5. Applications

A set of test samples from different origins including natural, commercial and restoration field samples was selected to test the feasibility of the SIMCA models and sub-models. The origin and kind of protein present in several of them were previously known, which made it possible to validate them. All the samples were treated as in the *Analytical Procedures* section; compositional amino acid profiles were obtained and their distances to the models/sub-models constructed subsequently calculated. Table 5 shows the SIMCA distances of each test sample to the models and sub-models established. Their variance values (s_i^2) were statistically compared to each model (s_o^2) by means of the F-test.

Four casein-like test samples were analyzed. Donkey and human colostrums milk-samples (1-3) initially collected as casein samples were outliers in their class and subsequently considered test samples in order to discover their nature. They were not classified as belonging to any class considered here, but it is remarkable that these samples showed a lower distance to the albumin class. This may mean that the protein present in these samples is like albumin. This fits with the fact that alpha-lactoalbumin is the principal protein in milk colostrums, even higher than casein whose content in colostrums is very low [43] (virtually nil). Only milk sample No. 4 is correctly identified as casein and barely as a cow casein (P=1%).

Table 5

Test samples 5-8 were called “only glue” because the only information available about them was that they were glues. Test samples 6 and 8 were correctly classified as collagens

(77 and 98 %, respectively) and subsequently as belonging to the mammal class with a high probability success (68 and 75 %, respectively); whereas 5 and 7 were not well classified as collagens (2 and 1 % probability success, respectively), perhaps because these test samples show a high content of Leu, Ser and Thr. Test samples 9-13 were all correctly assigned to the collagen class (P>70%) and subsequently identified as mammal samples. No identification of these samples according their animal part (bone, skin) was possible because the respective models could not be established. This is currently under study. Any protein identification strategy is subjected to a good selection of the appropriate standard reference set used. The advantage of the SIMCA method is that the identification of samples is performed only on the classes considered; making is possible to classify a sample as unknown or not belonging to any class.

Three artwork test samples from cultural heritage restoration works were considered (14-16). They were classified as fish collagen with a high confidence level.

6. Conclusions

We have elaborated a set of standard protein samples used as a reference to identify protein samples through their amino acid profile using the SIMCA pattern recognition technique. Thirteen SIMCA models at both class and subclass levels were developed and then optimized following variance and interclass distance criteria. We have improved the performance of SIMCA models respect to our previous approaches because of the use of interclass distance as optimization criteria and the increase in the number of available objects. These models were used to identify the binders in a set of test samples of different origins, showing the validity of models built. Successful identification was made possible

by the availability of various reference standards. The advantage of SIMCA is that the identification of the binder present in the samples is done only with classes that have been considered previously, making it possible to classify a sample as belonging to one of them or as unknown, i.e., not belonging to any class studied. Additionally, with the SIMCA method, it is possible to know what the proteinaceous binders are, not only at a class level but at a subclass level as well. This methodology can be applied to identifying the origin of protein binders in artworks in the field of conservation and restoration, may provide information about the historical provenance of materials in schools of art and might help to authenticate or refute questionable works of art. However, at this time it has one particular handicap: identifying mixtures of binders is difficult. New research is currently underway in this respect.

7. Acknowledgements

This study was supported by Research Group FQM118. The authors acknowledge financial support from the *Ministerio de Ciencia e Innovación, Dirección General de Investigación y Gestión del Plan Nacional de I+D+i* (Spain) (Project CTQ2009-14428-C02-01) and Junta de Andalucía (*Proyecto de Excelencia* P10-FQM-5974). These projects were partially supported by European Regional Development Funds (ERDF).

8. References

1 R. Aruga, P. Mirti, A. Casoli and G. Palla, *Fresenius' J. Anal. Chem.*, 1999***, 365*** (6), 559-566.

2 Pacheco, F. In *Arte De La Pintura*; Ed. Cátedra, Colección Arte: Madrid, Spain, 1990.

3 A. Karpowicz, *Stud. Conserv.*, 1981**, *26*** , 153-160.

4 E. Manzano, N. Navas, R. Checa-Moreno, L. Rodriguez-Simon and L. F. Capitan-Vallvey, *Talanta*, 2009**, *77*** (5), 1724-1731.

5 A. Nevin, I. Osticioli, D. Anglos, A. Burnstock, S. Cather and E. Castellucci, *Anal. Chem.*, 2007**, *79*** (16), 6143-6151.

6 M. P. Colombini, A. Andreotti, I. Bonaduce, F. Modugno and E. Ribechini, *Acc. Chem. Res.*, 2010**, *43*** (6), 715-727.

7 W. Fremout, J. Sanyova, S. Saverwyns, P. Vandenabeele and L. Moens, *Anal. Bioanal. Chem.*, 2009**, *393*** (8), 1991-1999.

8 G. Musumarra and M. Fichera, *Chemom. Intell. Lab. Syst.*, 1998**, *44*** (1,2), 363-372.

9 M. C. Gay, *Ann. Lab. Rech. Mus. France*, 1970**, *1*** , 8-24.

10 L. Masschelein-Kleiner, *PACT (Rixensart, Belg.)*, 1986**, *13*** , 185-207.

11 P. L. Jones, *Stud. Conserv.*, 1962**, *7*** , 10-16.

12 M. Johnson and E. Packard, *Stud. Conserv.*, 1971**, *16*** , 145-164.

13 L. S. Dolci, G. Sciutto, M. Guardigli, M. Rizzoli, S. Prati, R. Mazzeo and A. Roda, *Anal. Bioanal. Chem.*, 2008**, *392*** (1-2), 29-35.

14 A. Domenech-Carbo, F. B. Reig, J. V. G. Adelantado and V. P. Martinez, *Anal. Chim. Acta*, 1996**, *330*** (2-3), 207-215.

15 N. Navas, J. Romero-Pastor, E. Manzano and C. Cardell, *J. Raman Spectrosc.*, 2010**, *41*** (11), 1196-1203.

16 A. Nevin, I. Osticioli, D. Anglos, A. Burnstock, S. Cather and E. Castellucci, *J. Raman Spectrosc.*, 2008**, *39*** (8), 993-1000.

17 P. Vandenabeele, B. Wehling, L. Moens, H. Edwards, M. De Reu and G. Van Hooydonk, *Anal. Chim. Acta*, 2000**, *407*** (1-2), 261-274.

18 J. Romero-Pastor, C. Cardell, E. Manzano, A. Yebra-Rodriguez and N. Navas, *J. Raman Spectrosc.*, 2011**, *42*** (12), 2137-2142.

19 E. L. Ghisalberti and I. M. Godfrey, *Stud. Conserv.*, 1998**, *43*** (4), 215-230.

20 R. Checa-Moreno, E. Manzano, G. Miron and L. F. Capitan-Vallvey, *J. Sep. Sci.*, 2008**, *31*** (22), 3817-3828.

21 M. P. Colombini and F. Modugno, *J. Sep. Sci.*, 2004**, *27*** (3), 147-160.

22 M. T. Domenech-Carbo, *Anal. Chim. Acta*, 2008**, *621*** (2), 109-139.

23 A. Casoli, P. C. Musini and G. Palla, *J. Chromatogr. , A*, 1996**, *731*** (1 + 2), 237-246.

24 E. Kenndler, K. Schmidt-Beiwl, F. Mairinger and M. Pöhm, *Fresenius' J. Anal. Chem.*, 1992**, *342*** (1), 135-141.

25 I. Kaml, K. Vcelakova and E. Kenndler, *J. Sep. Sci.*, 2004**, *27*** (3), 161-166.

26 W. Fremout, M. Dhaenens, S. Saverwyns, J. Sanyova, P. Vandenabeele, D. Deforce and L. Moens, *Anal. Chim. Acta*, 2010**, *658*** (2), 156-162.

27 S. Kuckova, R. Hynek and M. Kodicek, *Anal. Bioanal. Chem.*, 2007**, *388*** (1), 201-206.

28 G. Leo, L. Cartechini, P. Pucci, A. Sgamellotti, G. Marino and L. Birolo, *Anal. Bioanal. Chem.*, 2009**, *395*** (7), 2269-2280.

29 C. Tokarski, E. Martin, C. Rolando and C. Cren-Olive, *Anal. Chem.*, 2006**, *78*** (5), 1494-1502.

30 A. Chambery, M. Di, C. Sanges, V. Severino, M. Tarantino, A. Lamberti, A. Parente and P. Arcari, *Anal. Bioanal. Chem.*, 2009**, *395*** (7), 2281-2291.

31 W. Fremout, S. Kuckova, M. Crhova, J. Sanyova, S. Saverwyns, R. Hynek, M. Kodicek, P. Vandenabeele and L. Moens, *Rapid Commun. Mass Spectrom.*, 2011**, *25*** (11), 1631-1640.

32 R. Pancella and R. Bart, *Z. Kunsttechnol. Konservierung*, 1989**, *3*** , 101-111.

33 Schilling, M. R.; Khanjian, H. P. James & James (Science Publishers eds.), London, 2011; pp 211-219.

34 M. P. Colombini, F. Modugno, M. Giacomelli and S. Francesconi, *J. Chromatogr. , A*, 1999**, *846*** (1 + 2), 113-124.

35 G. Gautier and M. P. Colombini, *Talanta*, 2007**, *73*** (1), 95-102.

36 R. Lleti, L. A. Sarabia, M. C. Ortiz, R. Todeschini and M. P. Colombini, *Analyst*, 2003**, *128*** (3), 281-286.

37 P. J. Gemperline, L. D. Webber and F. O. Cox, *Anal. Chem.*, 1989**, *61*** (2), 138-144.

38 R. Checa-Moreno, E. Manzano, G. Miron and L. F. Capitan-Vallvey, *Talanta*, 2008**, *75*** (3), 697-704.

39 S. Wold, *Pattern Recognition*, 1976**, *8*** (3), 127-139.

40 Cennini, C. In *The Craftsman's Handbook. Il Libro Dell'Arte*; Dover Publications, Inc.: New York, 1960.

41 J. E. Eastoe, *Biochem. J.*, 1957**, *65*** , 363-368.

42 M. Nemati, M. R. Oveisi, H. Abdollahi and O. Sabzevari, *J. Pharm. Biomed. Anal.*, 2004**, *34*** (3), 485-492.

43 L. Hambraeus, B. Lonnerdal, E. Forsum and M. Gebre-Medhin, *Acta Paediatr. Scand.*, 1978**, *67*** (5), 561-565.

Table 1. Collection of natural standard protein

ALBUMINS-Like poultry [(a)]	CASEINS-Like Milk samples of Spanish breeds	COLLAGENS-Like [(b)]
Chicken *Gallus gallus*	Caprine *Capra aegagrus hircus*	Fish
Dwarf chicken *Gallus gallus**	Granadina (3)	Flounder *Platichthys flesus* (3)
Pheasant *Phasianus colchicus* (2)	Granadino-Murciana	Cod *Gadus morruha* (5)
Goose *Anser anser*	Capra pyrenaica hispanica (2)	Sturgeon *Acipenser sturio* (8)
Turkey *Meleagris gallopavo* (2)	Malagueña (2)	Sole *Solea solea* (8)
Peacock *Pavo cristatus* (2)	Manchega (2)	Hake *Merluccius merluccius* (6)
Pigeon *Columba livia domestica*	Bovine *Bos Taurus*	Blue whiting *Micromesistius poutassou* (10)
Duck *Cairina moschata*	Friesian (2)	Turbot *Psetta maxima* (3)
Mallard *Anas platyrhynchos*	Holstein (2)	Mammalian:
	Brown Swiss	Rabbit *Oryctolagus cuniculus* (5)
	Jersey	Pigs *Sus* (5)
	Ovine *Ovis aries*	Bovine *Bos primigenius* (6)
	Segureña (2)	Ovine *Ovis aries* (8)
	Merino (3)	Caprine *Capra hircus* (8
	Red Majorcan	
	Manchega	
	Lacha (2)	
	Churro (2)	
	Castellana (2)	
	Awassi (2)	
	Black-eye	

[(a)] whites, yolks and whole eggs, [(b)] skins, backbones and air bladders of fish and skins, bones and cartilages of mammals

* *holland bredd* The number of replicated samples coming from different origins likes farms, supermarkets, etc. appears between brackets.

623

624

625

626

627

Table 2. Models of class and subclass

Albumin (N = 101, P = 9 A=2)																	
V: 78 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 0.77	Loading[1]:		0.40		0.36	0.19		0.40	0.37		0.24			0.37		0.28		-0.33
S: 96 %	Loading[2]:		0.11		0.25	0.52		0.12	0.08		0.49			0.21		0.48		-0.35
Sp: 100 %	MP:		0.54		0.52	0.49		0.58	0.45		0.53			0.50		0.62		0.54
Casein (N = 79, P = 10, A=2)																	
V: 73 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 0.80	Loading[1]:	0.34	0.42	-0.25		0.07	0.27	0.35	0.14		0.37		0.41			0.34		
S: 94 %	Loading[2]:	0.03	0.06	-0.44		0.53	0.27	0.08	0.51		0.28		0.05			0.33		
Sp: 100 %	MP:	0.36	0.67	0.50		0.42	0.31	0.39	0.44		0.63		0.60			0.57		
Collagen (N = 204, P = 9, A=2)																	
V: 75 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 0.82	Loading[1]:	0.17	0.18	0.38	0.40	0.42			0.38		0.36			0.37		0.20		
S: 92 %	Loading[2]:	0.62	0.59	-0.13	0.17	0.01			0.01		0.28			0.22		0.30		
Sp: 100 %	MP:	0.63	0.57	0.49	0.60	0.64			0.44		0.54			0.49		0.20		
	$DP_{albumin-casein}$:	11	21	7		5	11	19	14		11		6			7		
	$DP_{albumin-collagen}$:	4	13	37	23	65		1	21		15			10		35		6
	$DP_{casein-collagen}$:	16	69	41	34	94	8	22	23		11		7	25		26		
Glair (N = 33, P = 8, A = 2)																	
V: 80 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 0.81	Loading[1]:	0.45						0.44	0.41	0.47	0.29					0.10	0.34	-0.05
S: 97 %	Loading[2]:	0.01						0.25	0.23	0.10	0.36					0.56	0.37	0.54
Sp: 73 %	MP:	0.50						0.64	0.52	0.59	0.39					0.62	0.53	0.52
Whole egg (N = 32, P = 7, A = 2)																	
V: 78 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 0.88	Loading[1]:	0.15	0.27		0.32						0.52	0.46				0.44	0.35	
S: 97 %	Loading[2]:	0.65	0.50		0.47						0.09	0.05				0.30	0.07	
Sp: 84 %	MP:	0.71	0.5		0.53						0.79	0.48				0.62	0.23	
Yolk (N = 30, P = 8, A = 2)																	
V: 83 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 0.74	Loading[1]:					0.40		0.37	0.36		0.43		0.12		0.33	0.41	0.30	
S: 100 %	Loading[2]:					0.20		0.14	0.37		0.21		0.62		0.42	0.22	0.39	
Sp: 100 %	MP:					0.55		0.40	0.62		0.73		0.67		0.66	0.64	0.46	
	$DP_{glair-whole\ egg}$:	3.0	1.5		1.9			3.3	1.4	1.1	3.6	2.3				4.7	2.7	1.5

$DP_{glair-yolk}$:		3.7				7.4		7.9	6.3	1.7	5.6		4.2		11.0	9.0	7.5	4.6
$DP_{whole\ egg-yolk}$:		1.1	1.4		4.4	4.4		4.2	5.1		2.2	2.9	1.4		5.5	4.8	3.4	
Goat (N = 20, P =10, A =3)																		
V: 95 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 0.58	Loading[1]:	-0.41	0.27	0.40		-0.29		0.31			0.08	0.34	0.24	0.31			0.37	
S: 100 %	Loading[2]:	-0.13	-0.43	-0.15		-0.11		-0.38			0.49	-0.27	0.47	0.22			0.20	
Sp: 100 %	Loading[3]:	-0.16	0.18	-0.08		0.57		-0.18			-0.40	-0.28	-0.03	0.48			0.33	
	MP:	0.64	0.69	0.54		0.75		0.69			0.75	0.62	0.67	0.79			0.73	
Sheep (N =43, P =8, A =2)																		
V: 86 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 0.66	Loading[1]:	0.40	0.41					0.31		-0.29			-0.42			-0.37	0.24	0.36
S: 98 %	Loading[2]:	-0.11	-0.11					-0.47		-0.45			0.06			0.32	0.59	0.32
Sp: 81 %	MP:	0.58	0.63					0.65		0.53			0.67			0.67	0.67	0.58
Cow (N =15, P =11, A =4)																		
V: 98 %	Loading[1]:			0.29	-0.20	-0.03	0.28			-0.34		0.26	-0.39		-0.36	-0.32	0.32	0.36
Res. SD: 0.39	Loading[2]:			-0.31	0.45	0.50	0.35			-0.27		0.33	0.07		-0.13	0.24	-0.21	0.14
S: 93 %	Loading[3]:			-0.39	-0.02	-0.38	-0.17			-0.14		0.24	0.06		0.24	0.42	0.49	0.34
Sp: 100 %	Loading[4]:			0.01	0.17	-0.29	-0.30			0.18		0.77	0.01		-0.09	-0.27	-0.24	-0.19
	MP:			0.88	0.80	0.85	0.79			0.78		0.87	0.81		0.67	0.85	0.89	0.76
	$DP_{goat-sheep}$:	2.7	2.3	5.2		6.0		3.9		1.9	3.1	6.9	3.7	5.6		3.5	4.7	1.7
	$DP_{goat-cow}$:	4.7	1.4	6.2	9.8	6.7	10	4.9		4.8	3.4	8.4	5.8	2.5	7.9	4.8	5.4	4.2
	$DP_{sheep-cow}$:	2.9	1.5	9.5	7.3	6.0	8	2.2		3.5		7.7	1.6		6.7	1.4	2.3	3.6
Mammalian (N = 94, P =9, A =2)																		
V: 64 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 0.99	Loading[1]:	-0.44		0.43	-0.35			-0.04	0.14	0.23		0.29		0.41	0.41			
S: 92 %	Loading[2]:	-0.28		0.00	-0.09			0.53	0.46	0.45		-0.28		-0.30	-0.24			
Sp: 100 %	MP:	0.58		0.36	0.22			0.46	0.36	0.42		0.26		0.51	0.44			
Fish (N =121, P =7, A =2)																		
V: 67 %	Variables:	**Asp**	**Glu**	**HOpr**	**Ser**	**Gly**	**His**	**Arg**	**Thr**	**Ala**	**Pro**	**Tyr**	**Val**	**Met**	**Ile**	**Leu**	**Phe**	**L ys**
Res. SD: 1.04	Loading[1]:	-0.31	-0.29	0.11	0.23	0.49				-0.51				0.50				
S: 90 %	Loading[2]:	-0.48	-0.51	0.41	-0.48	-0.27				0.19				0.02				
Sp: 100 %	MP:	0.48	0.51	0.24	0.40	0.50				0.49				0.35				
	$DP_{mam-fish}$:	2.2	2	5	5.3	2.1		2.3	4.3	4.1		3.7		5.3	2.3			

V: Variance explained by the model, Res. SD: Residual standard deviation of the model, S: Sensitivity, Sp: Specificity, MP: modelling power, DP: Discriminant power

628

629

630

631

632

Table 3. Distance between subclasses

	albumin subclass			casein subclass			collagen subclass
Models:	G-W	G-Y	W-Y	C-O	C-V	O-V	M-F
All variables	1.5	2.9	1.8	1.7	2.3	1.4	1.4
Optimized	2.5	6.2	3.4	3.0	5.2	2.7	3.4

G: glair, W: whole egg, Y: yolk; C: caprine, O: ovine, V: vac

633

634

635

636

Table 4. Major class parameters for phylogenetic submodels.

Model	Variance explained	Objects	Variables	PCs	Res. SD	Sensibility	Specificity
Anseriformes	89 %	10	14	4	0.85	100 %	100 %
Galliformes & Columbiformes	73 %	27	13	3	0.82	89 %	85 %

	Asp	Glu	Ser	Gly	His	Arg	Thr	Ala	Tyr	Val	Met	Ile	Leu	Phe	Lys
Modelling power: Anseriformes:	0,62	0,52	0,61	0,63	0,79	-	0,26	0,40	0,35	0,61	0,71	0,62	0,65	0,72	0,64
Galliformes & Columbiformes:	-	0,54	0,32	0,58	-	0,60	0,58	0,62	0,60	0,49	0,25	0,38	0,49	0,55	0,58
Discrim. P. A/G&C:	7,9	6,4	6,6	3,2	3,6	0,5	8,9	5,5	2,7	3,1	3,5	4,2	3,1	4,3	2,0

Res. SD: Residual standard deviation of the class, Discrim. P.: Discriminant power

637

Table 5. Results of the SIMCA method. Distance (s_i^2) for the unknown samples to every class and subclass modelled. Critical distance to each model in brackets

	SIMCA distance to each class/subclass model												
			whole										
	alb	**glair**	**egg**	**yolk**	**anna**	**gallco**	**cas**	**cap**	**ov**	**cow**	**col**	**mam**	**fish**
Sample	**(0.77)**	**(0.81)**	**(0.88)**	**(0.74)**	**(0.85)**	**(0.82)**	**(0.80)**	**(0.58)**	**(0.66)**	**(0.39)**	**(0.82)**	**(0.99)**	**(1.04)**
1. Donkey milk	1,3	2,1	2,1	2,2	4,5	2,8	6,9	10,8	6,5	7,1	25	14	47
2. Human milk colostrum 1	1,2	2,3	1,9	2,1	4,4	3,6	6,7	12,5	7,4	9,3	24	17	44
3. Human milk colostrum 2	1,5	2,2	1,8	4,4	4,4	2,8	10,5	17,4	9,1	13,8	22	25	36
4. Cow milk	2,8	2,4	3,9	3,3	6,3	3,9	**0,63**	2,0	1,1	0,43	25	11	50
5. Skin glue	5,5	3,5	5,4	7,5	14	7,6	23	35	17	32	0,88	1,98	**0,61**
6. Strong glue	5,4	3,5	5,1	7,5	13	7,3	23	35	16	32	**0,44**	**0,57**	1,74
7. *Hausenblasen* Glue	5,4	3,3	5,3	7,4	13	7,5	23	34	17	32	0,94	2,47	**0,80**
8. Glue	5,4	3,4	5,1	7,5	13	7,3	23	35	16	32	**0,28**	**0,54**	1,74
9. Rabbit glue	5,5	3,5	5,3	7,5	14	7,4	23	35	16	32	**0,45**	**0,64**	1,69
10. Skin pork glue	5,4	3,5	5,2	7,5	13	7,4	23	35	16	32	**0,47**	**0,73**	1,94
11. Pork bone glue	5,4	3,5	5,1	7,4	13	7,4	23	35	16	32	**0,34**	**0,55**	1,58
12. Cow bone glue 1	5,5	3,6	5,3	7,6	14	7,5	23	35	16	32	**0,22**	**0,77**	1,42
13. Cow bone glue 2	5,4	3,5	5,2	7,5	13	7,4	23	35	16	32	**0,38**	**0,58**	1,61
14. Artwork sample 1	3,9	4,8	7,1	9,7	12	7.8	11	12	12	17	**0,79**	1,7	**0,71**
15. Artwork sample 2	4,4	4,7	7,5	9,3	14	7.6	11	14	12	17	**0,71**	1,4	**0,93**
16. Artwork sample 3	4,4	4,9	7,7	9,6	13	7.7	12	14	12	18	**0,65**	3,6	**0,65**

Probability level of belonging to class or subclass: (1-5 %) underlined, > 5% in **bold.**
Albumin (alb), annatidae (anna), galliformes and columbiformes (gallco), casein (cas), caprine (cap), ovine (ov), collagen (col), mammalian collagen (mam).

Figure captions

Figure 1. HPLC-UV chromatograms at 254 nm of 17 PTH-derivatives of amino acids presents in hydrolysates of representative binders: a) albumin, b) casein and c) collagen

Figure 2. Scores and loadings plot of the autoscaled chromatographic data of protein binder standards in the space of the first three principal components. Albumin (red), casein (green), and collagen (blue) classes.

Figure 3. Scores plot of subclasses a) albumins: (Y) yolk, (W) whole egg, (G) glair; b) caseins: (G) goat, (C) cow, (S) sheep; and c) collagens, (M) mammals, (F) fish.

Figure 4. Dendrogram built with 17 variables from the albumin class. Cluster 1 contains amino acids used in modelling whereas cluster 2 shows amino acids not used in the polished model.

Figure 5. Cooman's plot of the squared SIMCA distances obtained from the data set for Anseriformes (A) and Galliformes-Columbiformes (G) glair samples.

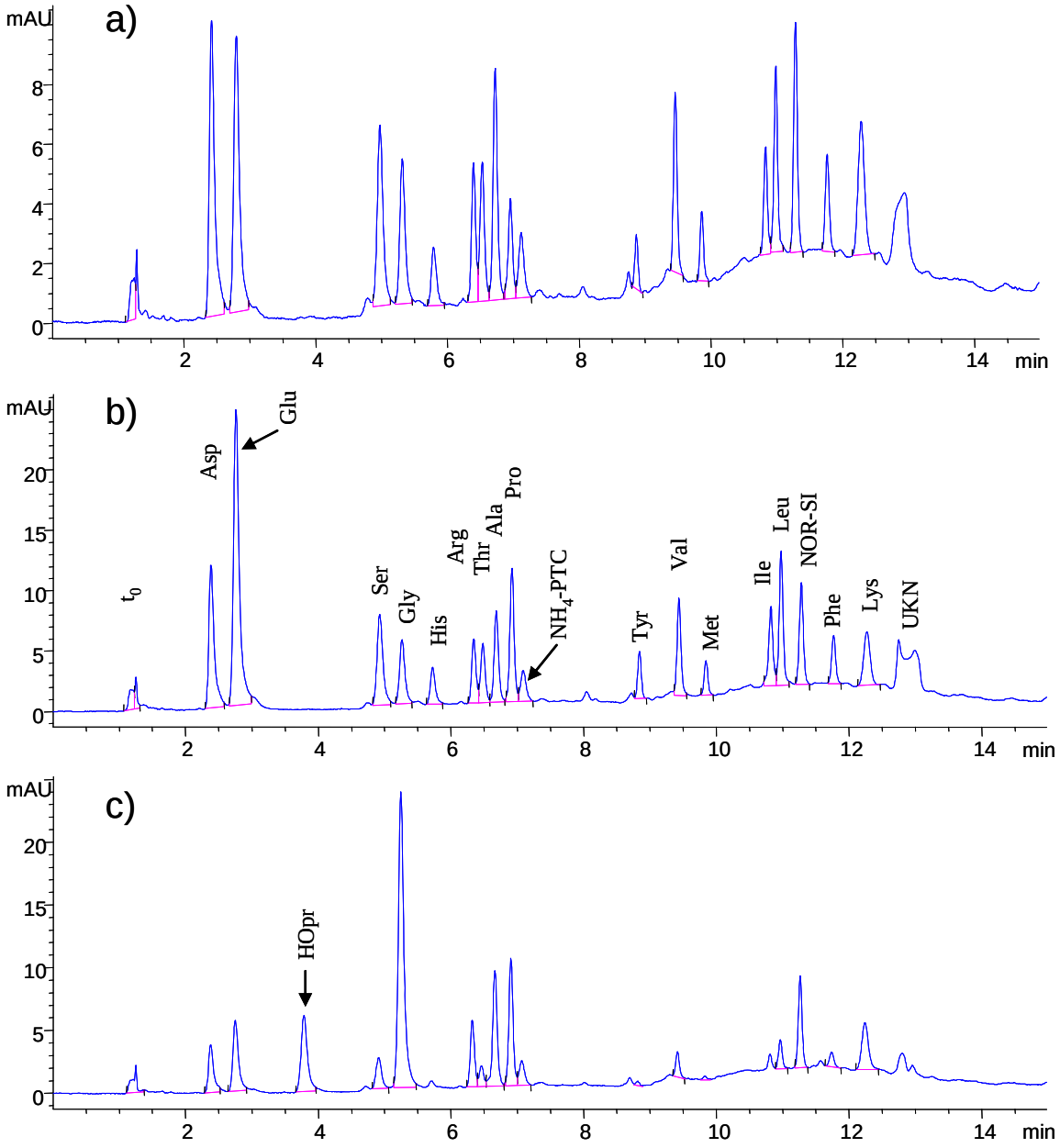


Figure 1

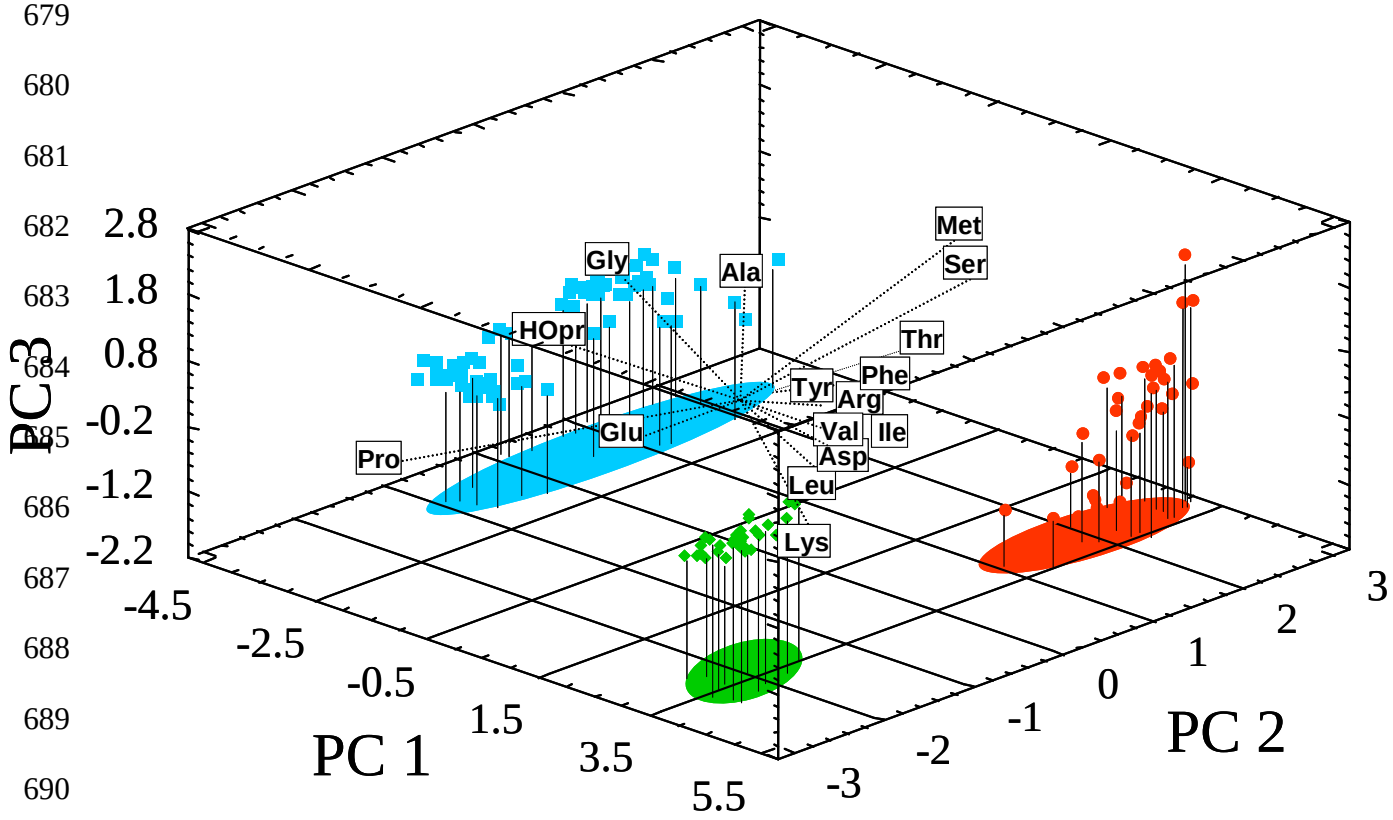


Figure 2

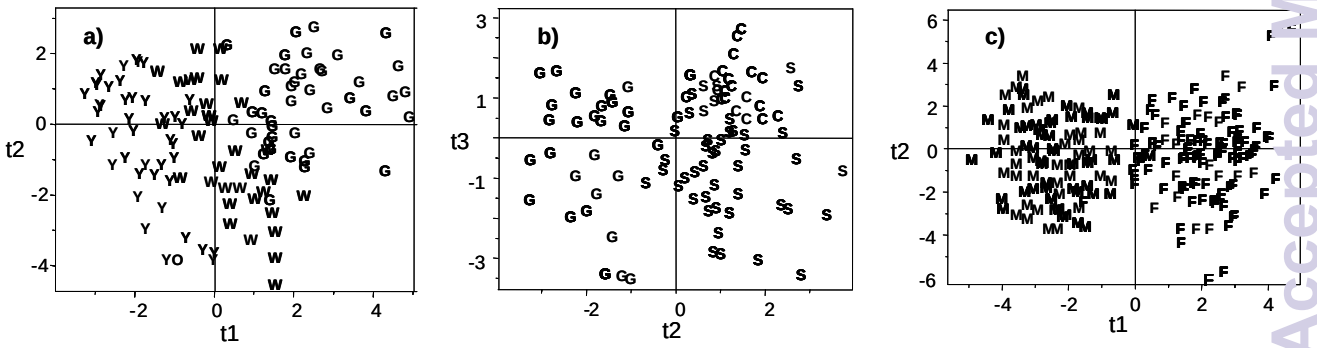


Figure 3

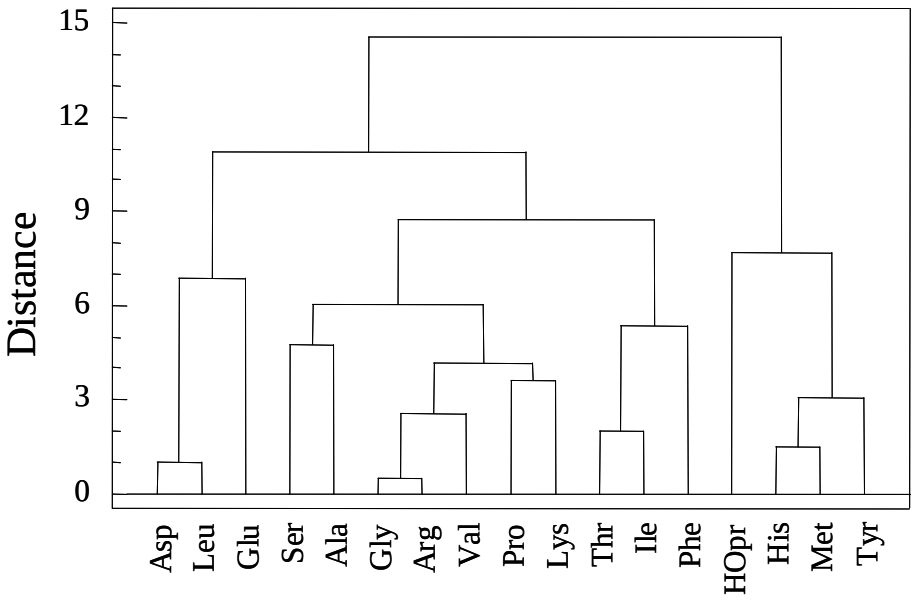


Figure 4

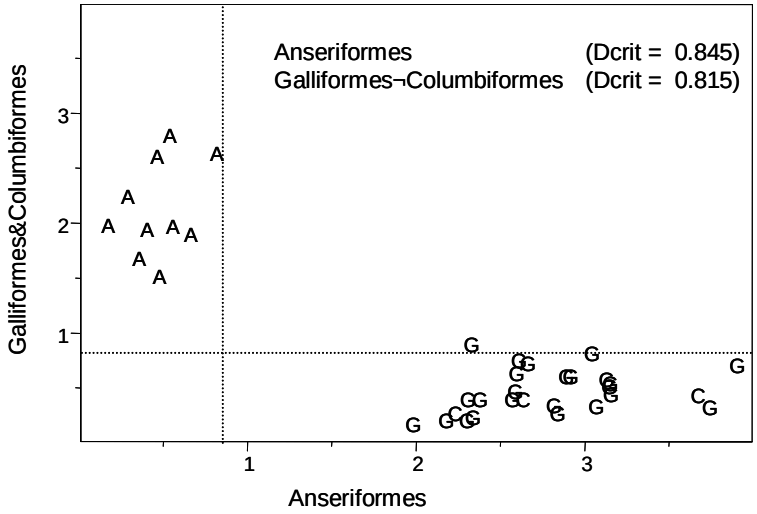


Figure 5