k NEAREST NEIGHBOURS ANALYSIS OF HUMAN SERUM CARBONYL PROTEINS USING CAPILLARY ELECTROPHOREGRAMS

A. VĂLEANU[#], MIHAELA ILIE, INES DIMA, CARMEN PURDEL

Toxicology Department, Faculty of Pharmacy, "Carol Davila" University of Medicine and Pharmacy, Bucharest, Romania, [#]e-mail: andrei_valeanu@yahoo.com

Abstract. Protein carbonyls are very important biomarkers of oxidative stress. Pattern recognition (PR) methods use certain mathematical algorithms to derive qualitative features of samples based on their parameters. The aim of the paper was to find a pattern in the electrophoregrams of carbonyl proteins from human serum albumin using PR techniques. Samples of previously oxidized human serum albumin (HA) were analyzed using capillary electrophoresis (CE) at 214 nm and 365 nm. Five samples were selected to build the pattern and several peak classes were created at the two selected wavelengths using cluster analysis. Several serum samples from diabetes patients were also analyzed and compared to those of the master samples using the k nearest neighbours method (kNN). The preliminary results indicate that a series of CE peaks can be found in most of the HA sample runs, with small variability in each peak class. Only a few of the patients' peaks were found to fit the pattern, with retention time being the most similar parameter. KNN and cluster analysis are both simple, yet efficient techniques, capable of a complex and profound analysis of the studied system and can be used to build the pattern of carbonylated human serum albumin.

Key words: protein carbonyls, capillary electrophoresis, pattern recognition, kNN algorithm, human serum albumin.

INTRODUCTION

Carbonyl proteins are amongst the most important biomarkers of oxidative stress. Their high levels and occurrence have broadly been associated with several pathologies, including neurodegenerative diseases (Alzheimer's, Parkinson's), metabolic syndrome, muscular dystrophy, rheumatoid arthritis, but also with abnormal physiological states (addiction, intoxication, etc.) [5, 11, 15]. Taking these aspects into account, it becomes of an utmost importance to find appropriate methods of analyzing and quantifying the protein oxidation process in such pathologies.

Received: January 2016;
in final form February 2016.

Many instrumental methods have been used so far for the detection and quantification of carbonyl proteins, especially absorbance or fluorescence spectrophotometry [8, 13], based on the reaction between 2,4-dinitrophenylhydrazine (DNPH) and the carbonyl groups of the analyzed protein. However, capillary electrophoresis techniques have rarely been used for such purposes [10]. It could therefore be of interest to develop such instrumental methods for building a reliable and reproductible pattern of the carbonylated human serum albumin. As generally raw data consist of a signal comprising a large amount of related numerical values (retention time, wavelength, absorption, area, etc.), an appropriate algorithm should be used to characterize each sample [6]. Chemometrics comes with the necessary tools for the enhanced characterization of large amounts of data. It implies using mathematical models that allow the conversion of the obtained stochastic data into relevant chemical information [2].

Pattern recognition is one of the most popular branches of chemometrics as it uses several mathematical algorithms to sort, reorganize, classify and/or group the instrumental data, in order to derive qualitative features of the samples. As the name suggests, the main objective of the technique is to discover regularity in the dataset since the raw results make impossible this type of detection without an appropriate mathematical analysis [2, 3].

Pattern recognition (PR) techniques can be divided into unsupervised and supervised methods. While supervised pattern recognition implies two datasets (the so-called training and test sets; the training set basically represents the “pattern” and the test set is used to check the features of the training set [2]), unsupervised PR, mainly cluster analysis, involves grouping raw, unknown data into certain classes (groups, clusters) based on the similarity between the obtained results [2, 9, 12, 19].

k-Nearest Neighbours (kNN) is one of the most applied supervised pattern recognition techniques, mainly used for the classification of unknown samples (test set) into one of the known classes (groups) that form the training set, based on a rank (k) representing the closest nearest objects to the unknown sample [8]. It has many applications in the pharmaceutical and chemical sciences, such as determination of geographical origins of plants, SAR and 3D-QSAR studies or classification of minerals [1, 4, 7, 17]. It was also used for the estimation of missing values of certain parameters defining a sample [18, 14], in two forms: IC-kNNI (Incomplete case – k-Nearest Neighbours Imputation) and CC-kNNI (Complete case – k-Nearest Neighbours Imputation). Both methods are widely used, however IC-kNNI offers a greater freedom in completing the missing values [16].

The number of the nearest neighbours (value of *k*) is generally a complex issue and its assignment must be made by carefully taking into account the size of the training set, the type of analysis which leads to the results, the proportion between the 2 datasets and also the number of training set classes. Other variables may also be taken into account [2].

The aim of the paper was to find a pattern in the electrophoregrams of carbonyl proteins from human serum albumin using PR techniques (Cluster Analysis, kNN classifier, IC-kNNI).

MATERIALS AND METHODS

SAMPLES

Human serum albumin (HA) for injection was bought from Baxter AG. Samples of about 5 mg protein/mL were subjected to an oxidation process using a 1 mM $FeCl_3$/25 mM potassium ascorbate mixture [10].

Serum samples from fasting venous blood of diabetes patients were separated via centrifugation.

Both HA and serum samples were derivatized with 2,4 dinitrophenyl-hydrazine (DNPH) and subjected to capillary electrophoresis in the following instrumental conditions: G1600A (Agilent) capillary electrophoresis system (CE) with diode-array detector (DAD) controlled by Agilent ChemStation ver. B.0 2.0× software; 64.5 cm (effective length 56 cm) × 75 μm i.d. fused-silica capillary column thermostated at 25^o C; 20 mM borate buffer/dextran 70 15% w/w, pH 9.0; hydrodynamic injection for 10s (0.5 psi, –25 kV); 35 min running time; detection at 370 nm, 365 and 214 nm [10].

PEAK SELECTION AND kNN CLASSIFIER

Only electrophoregrams acquired at 214 and 365 nm were analyzed using PR techniques. For this purpose, all the obtained electrophoretic peaks were manually integrated. A peak selection algorithm (cluster analysis method) was used for the electrophoregrams, in which only peaks with a height value greater than 2mAU (milli absorbance units) were selected and grouped. Considering these aspects, from the 7 analyzed HA samples, only 5 were chosen to build the pattern, the ones with the largest numbers of significant peaks.

Several peak classes (clusters) were formed using the retention time as a parameter for each run, at both wavelengths (214 and 365 nm), so that the normalised root of sum of squared differences (between each pair of peaks from one class), $NRSSD_{class}$ (Eq. 1), was less than 0.75 for the groups with more than two peaks and less than 0.5 for the classes with only two peaks (3 missing values, meaning that 3 runs did not present values greater than 2 mAU at the retention time corresponding to the chosen peak that defined the cluster). Additionally, the difference between each pair of average retention times/class had not to be smaller than 0.5, in order to avoid a too great similarity between the generated clusters. For a correct normalization, the sum of differences between the retention times was divided by the number of unordered peak pairs within the group.

$$\mathrm{NRSSD}_{\mathrm{class}} = \sqrt{\sum_{i,j=1}^{n} \frac{(t_i - t_j)^2 \times 2!(n-2)!}{n!}} \tag{1}$$

where t_i , t_j = the i^{th} and j^{th} retention time values from the peak class and n = the total number of retention time values from that class [2].

By the procedure described above, 13 peak classes were selected at 214 nm and 11 at 365 nm that constituted the training set (master samples – MS). Consequently, each peak class contains the peak retention time values for the 5 selected samples. However, there were several cases in which the peaks for the specified retention time missed.

In the next step, each electrophoretic peak obtained from the serum of the 6 diabetes patients was classified in one of the 13 (for 214 nm) and 11 (for 365 nm) peak classes, using the kNN technique and the Euclidean distance (ED) as a comparison parameter.

The kNN classifier implied several phases: (a) calculation of the ED between the unknown peak (t_p) and all the selected and grouped MS peaks, considering the retention time as parameter; (b) setting a rank (k) representing the most similar retention times to the retention time corresponding to the unknown test sample (TS) peak; (c) inclusion of the unknown peak in the class with the most members from the k nearest neighbours [2].

The ED was calculated using Eq. (2).

$$\mathrm{ED}_{p-m} = \sqrt{\left(t_p - t_m\right)^2} \tag{2}$$

where t_p is the retention time corresponding to the unknown TS peak, and t_m – the retention time corresponding to the MS peak [2].

A value of $k = 3$ was found to be optimal for the TS peaks classification. The choice was made due to the number of retention times from each peak class.

After the preliminary classification of the TS peaks in the MS classes, an algorithm was proposed in order to determine whether the classified TS peaks fit or not the carbonylated HA pattern. This involved also calculating the NRSSD between the patient peak and the peak class where it was included using the kNN method, according to Eq. (3).

$$\mathrm{NRSSD}_{\text{patient-class}} = \sqrt{\sum_{i=1}^{n} \frac{\left(t_p - t_i\right)}{n}} \tag{3}$$

where t_p is the retention time corresponding to the classified TS peak, t_i – the ith retention time from the the MS peak class and n is the number of retention times from the MS peak class.

In order to check if the peaks of the TS electrophoregrams fit the MS pattern, the following interpretation was made:

- If $\mathrm{NRSSD}_{\text{patient-class}} < \mathrm{NRSSD}_{\text{class}}$ on the basis of the retention time, the peak could fit the pattern;
- For each electrophoregram, the kNN classifier was applied to three parameters: retention time, peak height and peak area.
- If $\mathrm{NRSSD}_{\text{patient-class}} < \mathrm{NRSSD}_{\text{class}}$ on the basis of the peak retention time, area and height, the peak entirely fits the HA pattern.

IC-kNNI METHOD FOR THE ESTIMATION OF MISSING VALUES

For the IC-kNNI implementation, only peak classes with not more than two missing retention time values were chosen, in order to provide a realistic estimation. Hence, 7 remaining peak classes were selected for 214 nm and 6 for 365 nm.

The IC-kNNI technique is based on a simple principle: if a missing value exists for the hth peak class (variable) of the lth sample, the search should be performed within the samples with existing values for the hth variable and considering the highest number of common peaks with the lth sample. If only a single sample with the maximum number of common peaks exists, then its hth variable value will be inputed to the missing one of lth sample. If several samples with the maximum number of common peaks exist, then a similarity measure (e.g. ED) between these samples and the lth sample should be computed, afterwards the hth peak from the most similar sample should be chosen for input [16]**.**

This type of algorithm was applied for all the missing values estimation considering the electrophoregrams at both wavelengths (214 nm and 365 nm), and NED (Normalised Euclidean Distance) as numerical similarity parameter (Eq. 4). Only the retention time values were considered,

$$\mathrm{NED}_{al} = \sqrt{\frac{\sum_{j=1}^{m}\left(t_{aj} - t_{lj}\right)}{m}} \tag{4}$$

where *a* and *l* indicate the samples, *j* indicates the peak class, and *m* is the number of common peak classes between the two samples [2].

Last but not least, a conclusive analysis was performed in order to determine how the missing values input influences the NRSSD from each peak class and hence how useful this estimation might be in terms of variability reduction.

It was found more appropriate to select the input peak by analyzing the retention time values, because they reflect the exact moment when a certain compound was separated.

RESULTS AND DISCUSSION

PEAK SELECTION AND kNN CLASSIFIER

As mentioned above, no peak class with the complete number of 5 retention time values was found. Two peak classes presented 4 out of 5 retention time values at 214 nm, (which means that 4 samples out of 5 presented a retention time value for that specific peak), 5 classes contained 3 values and 6 groups presented only 2 retention times. A minimum $\text{NRSSD}_{\text{class}}$ of 0.0478 was found and a maximum of 0.5397.

Similar results were obtained for 365 nm: 1 peak class with 4 values, 5 classes with 3 retention times and 5 with only 2 values. A minimum $\text{NRSSD}_{\text{class}}$ of 0.0474 was determined and a maximum of 0.7248. Results showed that a normalised root of sum of squared differences within the class ($\text{NRSSD}_{\text{class}}$) value smaller than 0.75 ensures a good similarity and small variability between the retention times belonging to each class of peaks.

Table 1

Retention time values and NRSSDs for each peak class at 214 nm (44.61% missing values)

214 nm	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**$\text{NRSSD}_{\text{class}}$**
Class 1	4.1606	4.1295	4.4913			*0.2836*
Class 2	5.7599	5.7840		5.6394		*0.1096*
Class 3	7.8051	7.2438		7.3345	6.8899	*0.5331*
Class 4	8.4272	8.1865			8.2897	*0.1708*
Class 5	9.9307				9.8829	0.0478
Class 6	11.5201		11.9963	11.9557	11.6812	*0.3211*
Class 7	12.4177		13.1238		13.0218	*0.5397*
Class 8			14.2651		14.2065	0.0586
Class 9	15.7219			15.3300		0.3919
Class 10		19.8532	19.9994			0.1462
Class 11	21.1809			21.4630	21.5127	*0.2531*
Class 12	28.8867			28.4217		0.4650
Class 13				30.3221	30.6855	0.3634

The data contained in these peak classes was considered to form the pattern of carbonylated human serum albumin. The detailed results are presented in Table 1 (214 nm) and Table 2 (365 nm). The NRSSD values for the classes with more than 2 retention times are written in italics.

Table 2

Retention time values and NRSSDs for each peak class at 365 nm (47.27% missing values)

365 nm	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**NRSSD$_{\text{class}}$**
Class 1	4.1615	4.1276	4.4901			*0.2832*
Class 2	5.5858	5.7826		5.6383		*0.1441*
Class 3	7.8058	7.7474			6.8904	*0.7248*
Class 4	8.4166	8.1861			8.2882	*0.1633*
Class 5	9.9283				9.8809	0.0474
Class 6	11.3407				11.0266	0.3141
Class 7	11.5185		11.9972	11.9540	11.8003	*0.3062*
Class 8	12.4225		13.1232		13.0214	*0.5354*
Class 9			14.2625		14.2062	0.0563
Class 10	21.1842				21.5107	0.3265
Class 11	28.8819			28.4201		0.4618

After the peak selection process, all the peaks obtained from the 6 TS were succesfully included in one of the 13 or 11 peak classes at a *k* value of 3 (for 214 nm and 365 nm, respectively). A total number of 73 TS peaks were classified at 214 nm and 42 at 365 nm.

Only a few TS peaks were found to entirely fit the pattern (5 at 214 nm and 4 at 365 nm). Others corresponded to the carbonylated HSA pattern by considering only the retention time (4 at 214 nm and 1 at 365 nm). The retention time was the most similar parameter, proving more qualitative and less quantitative resemblance. In this case, the retention time superposition was considered a necessary, but not sufficient condition; thereby, the resemblance obtained by considering only the peak area and height is irrelevant. The lack of pattern superposition might be due to the fact that the samples obtained from the serum of the diabetes patients have a more complex composition and contain other proteins besides the HA and/or other components that react with DNPH.

IC-kNNI RESULTS

The results of the IC-kNNI analysis are presented in Table 3 (for 214 nm) and Table 4 (for 365 nm). The inputs marked with $(^*)$ are the ones in which no similarity determination was necessary, due to the fact that only one sample with the maximum number of common peaks existed. The other cases are marked with

($^{\#}$). Generally, after the input, NRSSD$_{\text{class}}$ decreased, hence the variability in the peak classes was diminished. However, we found a situation where the input led to a NRSSD$_{\text{class}}$ increase (marked with italics), but the difference of 0.0022 was considered to be insignificant.

Table 3

IC-kNNI results for 214 nm and retention time parameter (34.28% missing values)

214 nm	**S1**	**S2**	**S3**	**S4**	**S5**	**NRSSD before IC-kNNI**	**NRSSD after IC-kNNI**
C1	4.1606	4.1295	4.4913	4.1606*	4.1606*	0.2836	0.2149
C2	5.7599	5.7840	5.7599*	5.6394	5.7599*	0.1096	0.0814
C3	7.8051	7.2438	6.8899$^{\#}$	7.3345	6.8899	*0.5331*	*0.5353*
C4	8.4272	8.1865	8.4272*	8.2897$^{\#}$	8.2897	0.1708	0.1459
C5	11.5202	11.9557$^{\#}$	11.9963	11.9557	11.6812	0.3211	0.2975
C6	12.4177	13.1238$^{\#}$	13.1238	13.1238$^{\#}$	13.0218	0.5397	0.4350
C7	21.1809	21.1809*	21.1809*	21.4630	21.5127	0.2531	0.2391

S = sample, C = Class

Table 4

IC-kNNI results for 365 nm (36.66% missing values)

214 nm	**S1**	**S2**	**S3**	**S4**	**S5**	**NRSSD before IC-kNNI**	**NRSSD after IC-kNNI**
C1	4.1615	4.1276	4.4901	4.4901$^{\#}$	4.1615*	0.2832	0.2640
C2	5.5858	5.7826	5.5858*	5.6383	5.5858*	0.1441	0.1205
C3	7.8058	7.7474	6.8904$^{\#}$	6.8904$^{\#}$	6.8904	0.7248	0.6871
C4	8.4166	8.1861	8.4166*	8.4166$^{\#}$	8.2882	0.1633	0.1481
C5	11.5185	11.5185$^{\#}$	11.9972	11.9540	11.8003	0.3062	0.3257
C6	12.4225	12.4225$^{\#}$	13.1232	13.1232$^{\#}$	13.0214	0.5354	0.5198

S = sample, C = Class

The number of peak classes was reduced in order to improve the missing values percentage and hence to provide a more realistic value estimation.

CONCLUSIONS

Peaks of the capillary electrophoregrams can be used to build a pattern of the oxidized human serum albumin. The pattern was built using the cluster analysis and *k*-nearest neighbour (kNN) algorithms, using the retention time as variable and the Euclidean Distance as a similarity parameter. The normalised root of sum of squared differences used as numerical indicator assured a good similarity inside each class of peaks. The missing values percentage was enhanced by the careful selection of the peak classes which were subjected to the IC-kNNI data estimation algorithm. From the 13 peak classes (7 at 214 nm and 6 at 365 nm) on which IC-kNNI was applied, only one presented an increase in the $\mathrm{NRSSD}_{\mathrm{class}}$ value. Taking these aspects into consideration, this missing values estimation method can be applied in the cluster analysis technique, as an effective tool for similarity enhancement within each group (cluster).

The combination of the two types of kNN algorithms (kNN classifier and IC-kNNI), within the same analysis reflects the complexity and efficiency of this kind of data analysis techniques. Further studies could combine kNN with other pattern recognition methods and also develop more complexly associated mathematical algorithms. This new generated models could also be used to compare electrophoregrams of the patients with the pattern obtained for *in vitro* oxidized samples, after the missing values input.

Acknowledgements. This paper was co-financed from the European Social Fund, through the Sectorial Operational Programme Human Resources Development 2007–2013, contract POSDRU/159/1.5/S/155463 "Supporting excellence in scientific interdisciplinary doctoral research in the economic, medical and social fields", coordinator The Bucharest University of Economic Studies.

REFERENCES

1. BHADORIYA, K.S., M.C. SHARMA, S. SHARMA, S.V. JAIN, M.H. AVCHAR, An approach to design potent anti-Alzheimer’s agents by 3D-QSAR studies on fused 5,6-bicyclic heterocycles asc-secretase modulators using kNN-MFA methodology, *Arabian Journal of Chemistry,* 2014, **7**, 924–935.
2. BRERETON, R.G., *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, pub. Wiley, Chichester, 2003.
3. BRERETON, R. G., Pattern recognition in chemometrics, *Chemometrics and Intelligent Laboratory Systems,* 2015, **149**, 90–96.
4. CAO, D.S. *et al*., Kernel k-nearest neighbor algorithm as a flexible SAR modeling tool, *Chemometrics and Intelligent Laboratory Systems*, 2012, **114**, 19–23.
5. DEAVALL, D.G., E.A. MARTIN, J.M. HORNER, R. ROBERTS, Drug-Induced Oxidative Stress and Toxicity, *Journal of Toxicology*, 2012, Article ID 645460, doi:10.1155/2012/645460.
6. LAUER, H.H., G.P. ROZING, High Performance Capillary Electrophoresis. A Primer, *Agilent Tehnologies Inc.*, 2014, Publication Number 5990–3777EN.
7. LI, B., Y. WEI, H.G. DUAN, L. XI, X. WU, Discrimination of the geographical origin of *Codonopsis pilosula* using near infrared diffuse reflection spectroscopy coupled with random forests and k-nearest neighbor methods, *Vibrational Spectroscopy*, 2012, **62**, 17–22.
8. MESQUITA, C.S., R. OLIVEIRA, F. BENTO, D. GERALDO, J.V. RODRIGUES, J.C. MARCOS, Simplified 2,4-dinitrophenylhydrazine spectrophotometric assay for quantification of carbonyls in oxidized proteins, *Analytical Biochemistry*, 2014, **458**, 69–71.
9. MORLOCKA, G.E., P. RISTIVOJEVIC, E.S. CHERNETSOVA, Combined multivariate data analysis of high-performance thin-layer chromatography fingerprints and direct analysis in real time mass spectra for profiling of natural products like propolis, *Journal of Chromatography A,* 2014, **1328**, 104–112.

10. PURDEL, C., I. DIMA, D. MARGINĂ, D. GRĂDINARU, M. ILIE, Investigation of the carbonylation process of protein induced "in vitro" by different hydroxyl radical generating systems, *Revista de Chimie Buc.,* 2015, **66(3)**, 320–333.
11. ROGOWSKA-WRZESINSKA, A., K. WOJDYLA, O. NEDIC, C. P. BARON, H. R. GRIFFITH, Analysis of protein carbonylation – pitfalls and promise in commonly used methods, *Free Radical Research,* 2014, **48(10)**, 1145–1162.
12. SAURINA, J., Characterization of wines using compositional profiles and chemometrics**,** *Trends in Analytical Chemistry,* 2010, **29(3)**, 234–245.
13. STOCKER, P., E. RICQUEBOURG, N. VIDAL, C. VILLARD, D. LAFITTE, L. SELLAMI, S. PIETRI, Fluorimetric screening assay for protein carbonyl evaluation in biological samples, *Analytical Biochemistry*, 2015, **482**, 55–61.
14. TODESCHINI, R., Weighted k-nearest neighbour method for the calculation of missing values, *Chemometrics and Intelligent Laboratory Systems,* 1990, **9(2)**, 201–205.
15. UTTARA, B., V. SINGH, P. ZAMBONI, R.T. MAHAJAN, Oxidative stress and neurodegenerative diseases: a review of upstream and downstream antioxidant therapeutic options*, Current neuropharmacology*, 2009, **7(1)**, 65–74.
16. VAN HULSE, J., T.M. KHOSHGOFTAAR, Incomplete-case nearest neighbor imputation in software measurement data*, Information Sciences*, 2014, **259**, 596–610.
17. VARMUZA, K., P. FILZMOSER, M. HILCHENBACH, H. KRÜGER, J. SILÉN, KNN classification – evaluated by repeated double cross validation: Recognition of minerals relevant for comet dust, *Chemometrics and Intelligent Laboratory Systems,* 2014, **138**, 64–71.
18. WASITO, I., B. MIRKIN, Nearest neighbour approach in the least-squares data imputation algorithms, *Information Sciences*, 2005, **169**, 1–25.
19. YÜCEL, Y., P. SULTANOĞLU, Characterization of Hatay honeys according to their multi-element analysis using ICP-OES combined with chemometrics, *Food Chemistry,* 2013, **140**, 231–237.