Teaching Chemometrics with a Bioprocess: Analytical Methods Comparison Using Bivariate Linear Regression

Vanina G. Franco,[†] Victor E. Mantovani,[†] Hector C. Goicoechea,[†,*] and Alejandro C. Olivieri[‡,*]

Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, Ciudad Universitaria, Santa Fe S3000 CC. 242, Argentina, hgoico@fbcb.unl.edu.ar; Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, Rosario S2002LRL, Argentina, aolivier@fbioyf.unr.edu.ar

Received April 30, 2002. Accepted August 5, 2002.

Abstract: We present an advanced analytical chemistry laboratory experiment involving chemometrics. Students perform a comparison of two analytical methods by checking several analyte concentrations within a certain range by using least-squares linear regression. They obtain statistical information such as the presence of constant and proportional biases. The exercise is based on the determination of glucose levels using two colorimetric methods (enzymatic and Somogyi–Nelson) in a very simple batch system formed by an infusion of tea, glucose, and a combination of a yeast (*Schizosacaromyces pombe*) and a bacteria (*Acetobacter xylimun*), usually named *Kombucha*. Several samples are collected during a week of laboratory work, and measurements are performed in a subsequent four-hour laboratory class. Although commercial computer software exists for a variety of statistical applications, specific programs for the application of statistics to analytical chemistry are not prevalent. In order to solve this particular problem, a Matlab 5.3 routine is presented.

Introduction

The discipline of chemometrics studies the production of data and the extraction of information from them. A usual chemometric topic is method validation, which is devoted to ensuring the quality of an analytical method. This is important because if the qualities of the measurement process and of the produced data are not good enough, the chemical information may be uncertain or even wrong [1]. One of the most important features of an analytical method is its accuracy, and the usual way in which accuracy is assessed is by comparison of the concentration values obtained by the proposed method with those provided by a reference method [2, 3]. It is common practice to carry out this procedure by checking several analyte concentrations within a certain range and to perform a least-squares linear regression to give statistical information, such as the presence of constant and proportional biases. Several statistical tests comparing the intercept and slope values obtained by linear regression with the theoretical values (0 and 1, respectively) have been proposed in the literature [3–5]. The best approach seems to take into account the errors in both coordinate axes and the covariance between the slope and the intercept, which can be done by applying a joint confidence test for the intercept and the slope [6].

In order to develop this kind of statistical test, one may study different samples spanning a wide concentration range. A particular experiment presenting this characteristic is a bioprocess. It can be described as a process in which microorganisms convert chemical substrates into products of a higher value. These products may be of vital importance to modern society, ranging from traditional products, such as beer, to fine chemicals, such as antibiotics [7]. If one obtains the concentration values of a particular analyte at different times by applying a tested method and a reference one, a least-squares fit may be employed to compare the relative accuracy of the new method.

In the present work, students analyze the glucose fermentative degradation produced by a combination of a yeast (*Schizosacaromyces pombe*) and a bacteria (*Acetobacter xylimun*), usually named *Kombucha*, which constitutes a generous probiotic producer [8, 9]. The exercise is based on the determination of the glucose concentration levels by using two colorimetric methods (enzymatic and Somogyi–Nelson) in a very simple batch system formed by an infusion of tea, glucose, and *Kombucha*. The proposed system is especially suitable for undergraduate students for a number of reasons, namely: (1) the selected micro-organisms are not pathogenic, are safe, easy to manipulate, and resistant to contamination; (2) a large number of samples can be obtained in a working period of about one week; and (3) it offers the possibility of performing direct sampling and quantitative determination in the course of the process. The determinations are relatively rapid and can be completed during a laboratory period (approximately four hours) during the next week of class. It is an interdisciplinary work in which students obtain a biotechnology product and monitor one of the consumed substances. The results allow students to gain insight into the characteristics of the process and to carry out real-world chemical applications in analytical chemistry laboratories. Although commercial computer software exists for a variety of statistical applications, specific programs for the application of statistics to analytical chemistry are not prevalent. In particular, method comparison considering errors in both coordinate axes is not a method usually available in commercial software.

* Address correspondence to this author.
† Universidad Nacional del Litoral, Ciudad Universitaria
‡ Universidad Nacional de Rosario

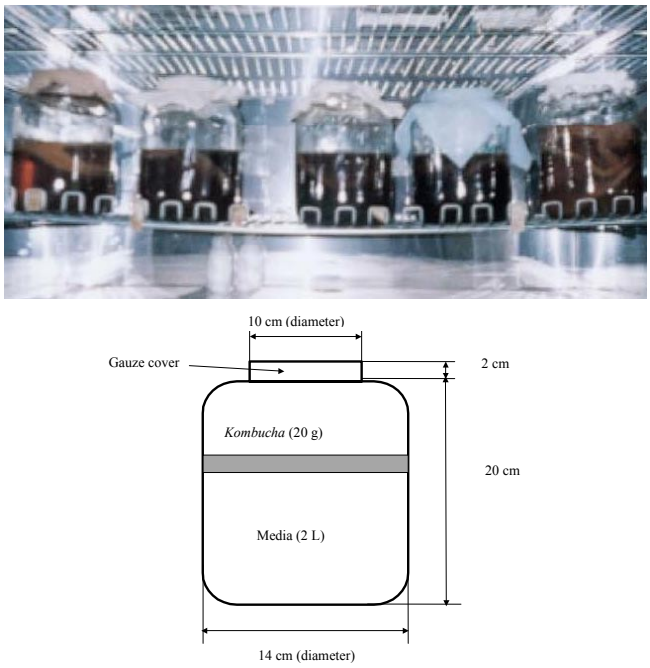

Figure 1. The batch reactor setup used in the glucose fermentation process in an infusion where Kombucha was added.

Experimental

Batch Reactor Setup. An infusion of commercial black tea and glucose was used as the medium, and a 3-L cylindrical vessel of thermal glass (3-L culture flask: Cole Palmer, Catalog Number E-29300-04) was used as the reactor (Figure 1). It can be set up in the laboratory using a 3-L-capacity glass flask obtained at any bazaar and a piece of gauze. After sterilization at 105 ºC, the reactor was charged with 2 L of medium to which 20.0 g of Kombucha wet mass was added. Then, the system was stored at 37 ºC.

Apparatus. Electronic absorption measurements were carried out on a Perkin–Elmer Lambda-20 spectrophotometer, using 1.00-cm quartz cells. All data were transferred to a PC Pentium 550 microcomputer for subsequent manipulation. An in-house Matlab 5.3 routine was used for statistical analysis and treatment of data.

Reagents. Analytical-reagent-grade chemicals and distilled water were used in all experiments. A black tea infusion (Camellia sinensis, Cameliaceas) was used as the medium. It was prepared with 2.5 g L^{-1} of commercial black tea and 65 g L^{-1} of glucose [50-99-7] dissolved in mineral water. It was then cooled to 28 ºC (or ambient temperature), and the pH was adjusted to 4.0 with a solution of acetic acid (1 mol L^{-1}).

Kombucha can be obtained at Kombucha Magic Mushroom Farm Inc., P.O. Box 20717, Cherokee Station, New York, NY 10021-0074, price: U.S. $50.00 (initial culture) or F. J. Perron, Natural Kombucha Cultures, Box 578, Lively, Ontario P3Y 1145, Canada, price: U.S. $35.00 (initial culture).

Medium preparation. (a) Heat the water; (b) add the glucose and dissolve; (c) heat to the boiling point, remove from heat, add tea, and let settle; (d) filter the tea and cool to ambient temperature; (e) pour into the glass container and add Kombucha; and (f) keep the flask in a cool airy place. Caution: do not expose to sunlight, do not smoke in the room, keep Kombucha in liquid, and do not move the culture.

Sampling. During the 5-day process, students or instructors obtained samples every 5 hr, although regular periods are not necessary. Each sample (ca. 20 mL of culture) was partitioned into a number equal to the laboratory groups and then preserved in the freezer at –20 ºC. The students were distributed among the groups (one for each of the samples to be taken).

Glucose determination. The Trinder enzymatic method was used as the reference technique to determine glucose using a diagnostic kit (colorimetric SIGMA kit, catalog No. 315-100) [10–12]. The Somogyi–Nelson procedure [11, 13] was the test method. Three replicates of the diluted sample were analyzed. (Dilution was required in order to bring them into the dynamic range of the calibration curve.)

The color reactions were carried out in the sample, blank (distilled water), and standard samples in order to construct calibration curves. Replicates were diluted 25 or 50 times in order to bring them into the dynamic range. The volumes for the Trinder color reaction were 20.0 μL for the test solution and 2.0 mL for the color reaction; the absorbances were read at 505 nm, and the selected standard concentrations were 5, 10, 15, 20, and 25 × 10^{-3} mol L^{-1}. These concentrations lie within the linear absorbance concentration range. The initial glucose concentration was ca. 3.5 × 10^{-1} mol L^{-1}; thus, the first sample (at zero time) was diluted 50 times.

For the Somogyi–Nelson color reaction the following reagents were prepared.

(a) two reagents to precipitate proteins: 0.06 mol L^{-1} sodium hydroxide and 10 g L^{-1} zinc sulphate;
(b) cupric reagent: to 28 g disodium phosphate and 40 g potassium sodium tartrate in 700 mL of distilled water was added 100 mL of 1 mol L^{-1} sodium hydroxide, 80 mL 100 g L^{-1} cupric sulphate, and 180 g of sodium sulphate; then, the solution was brought to 1000 mL and filtered;
(c) molybdic–arsenic reagent: 25 g sodium molybdate was dissolved in 450 mL of distilled water, 21 mL of sulphuric acid concentrate and 25 mL of a solution containing 3 g of sodium arsenate heptahydrate were added, and the solution was kept at 37 ºC for 24 hours.

The volumes used for the Somogyi–Nelson method were 100.0 μL for the test solution and 0.95 mL of 0.06 mol L^{-1} sodium hydroxide, and 10 g L^{-1} zinc sulphate. Then, the solution was centrifuged. Subsequently, 0.50 mL of supernatant was taken and mixed with 1.0 mL of reagent b (above). The solution was boiled for 15 min, cooled, and 1.0 mL of reagent c (above) and 5 mL of distilled water were added. Finally, the absorbances were read at 530 nm. The selected standard concentrations and initial glucose concentration were similar to those used for Trinder method.

General procedure. Sampling was performed during the first week of the experiment and measurements were taken during the second week. A week after the culture development, each group of students analyzed 16 different samples with 3 replicates each. With all data collected throughout the process, students performed the corresponding statistical test to know if the studied method presents bias.

Details of the purchasing, setup, routine, and experiment are located in the supporting material.

Theoretical Background

Linear regression is often employed for method comparison. The idea is to regress the concentration values predicted by a tested method against those provided by a reference analytical method. Once the slope and intercept of the fitted line are obtained, a test is applied in order to assess whether they are statistically equal or different than 1 and 0, respectively.

Linear regression is often performed nonweighted, that is, by giving equal weight to all the experimental data points. This is the so-called ordinary least-squares (OLS) procedure, in which the slope and intercept of the regression line is found by minimizing the following parameter,

$$U = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \quad (1)$$

where $\hat{y}_i$ is the value estimated by the regression line (according to the equation $y = a + bx$). The equations giving the values for the estimated slope ($\hat{b}$) and intercept ($\hat{a}$) are:

$$\hat{b} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \quad (2)$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad (3)$$

where $\hat{x}$ and $\hat{y}$ are the mean values of the x and y data, respectively, and N is the number of data points.

If points lying on the y axis are subjected to a variance significantly different than those on the x axis, then it would be wise to weight the y values as inversely proportional to their variances. This constitutes the popular weighted-least-squares (WLS) regression analysis, in which U is defined as

$$U = \sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)^2}{s_{yi}^2} \quad (4)$$

where s_{yi}^2 is the variance computed for each y_i value. In this case, the equations for the estimated slope and intercept are

$$\hat{b} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) / s_{yi}^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2 / s_{yi}^2} \quad (5)$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad (6)$$

In the event that both y and x values are affected by random errors of comparable magnitude, then none of the above approaches can be applied. In this case, it has been suggested that the slope and intercept should be found by minimizing the following parameter,

$$U = \sum_{i=1}^{N} \left[\frac{(\hat{y}_i - y_i)^2}{s_{yi}^2 + b^2 s_{xi}^2} \right] \quad (7)$$

where s_{xi}^2 represents the variance for each x_i data point. No explicit equation exists for solving the above problem. Instead, the recommended procedure is to find the slope and intercept iteratively. This method is called bivariate least-squares regression (BLS).

Once the slope and intercept are found by any of the above least-squares methods, the comparison of the latter values with the ideal ones (0 and 1 respectively) should be carried out by calculating the elliptic joint confidence region (EJCR) for both parameters. The equation describing the joint region is

$$N(a - \hat{a})^2 + 2\left(\sum x\right)(a - \hat{a})(b - \hat{b}) + \left(\sum x\right)^2 (b - \hat{b})^2 = 2s^2 F_{\alpha,2,N-2} \quad (8)$$

where N is the number of data points, s^2 is the regression variance, and $F_{\alpha,2,N-2}$ is the statistical F value with 2 and $N - 2$ degrees of freedom at a given $100 \times (1-\alpha)$ confidence level, usually 95%. If the point (1, 0) is inside the EJCR, it can be concluded that constant and proportional biases are absent.

Results and Discussion

In order to compare analytical methods by means of a linear regression, it is necessary to have a significant number of samples of various concentrations of the analyte under study. This makes the bioprocess discussed in this article an interesting system for this type of analysis, because both substrates as well as the fermentation products vary in concentration during the development of the experiment. In the specific case of *Kombucha*, glucose is consumed while organic acids are the main products. The glucose concentration starts at a value of 65 g L^{-1} and may end at values of 20 g L^{-1} in a period of five days [9]. If samples are taken at different times during the fermentation, a large number of samples will be available for the application of regression methods in order to compare two analytical methods.

Glucose is an interesting target in this regard because a number of simple and rapid analytical methods, which can be performed within a laboratory class (estimated in ca. 4 hr), exist for its determination. Among the available methods, the enzymatic method has proven efficiency, high accuracy, reproducibility, and selectivity, and it can be considered a reference method against which other less efficient methods can be compared. We selected the colorimentric Somogyi–Nelson method as the one to be tested.

During the first week of laboratory work, sixteen samples, which provided different glucose levels, were collected, corresponding to different steps in the bioprocess. They were preserved until the next week in a freezer at –20 °C. Each group of students defrosted their samples and carried out measurements applying both the tested and the reference method in triplicate for each sample.

Table 1 shows the data obtained when analyzing these sixteen samples by applying both methods. The corresponding averages and standard deviations are shown in Table 2. As can be seen, important differences exist between the variance corresponding to different concentration values. Figure 2 shows a comparative view of these standard deviations. It can be clearly seen that the standard deviations corresponding to the Somogyi–Nelson method are larger than the ones computed using the enzymatic method.

These values were subjected to the OLS, WLS, and BLS techniques and the results obtained by these methods are shown in Table 3. As can be seen, different conclusions may be obtained when the statistical technique is not properly chosen. Figure 3 shows the ideal line that corresponds to the hypothetical parameters $a = 0$ and $b = 1$ and that corresponding to different adjustments.

Table 1. Concentration Values Obtained by Applying the Enzymatic (E) and the Somogyi–Nelson (S–N) Methods (# is sample number; values are glucose in g L^{-1})

	Method			Method			Method	
#	E	S–N	#	E	S–N	#	E	S–N
1	13.63	21.45	6	26.98	41.70	11	40.51	44.03
1	13.72	23.24	6	27.25	43.85	12	40.87	48.51
1	13.81	22.88	7	31.07	42.96	12	41.06	48.33
2	23.80	38.12	7	31.34	41.34	12	41.15	48.87
2	23.89	37.76	7	31.34	44.57	13	43.88	49.77
2	23.98	32.74	8	31.70	45.65	13	43.96	50.13
3	25.16	37.40	8	31.79	44.57	13	43.99	49.95
3	25.34	38.65	8	31.88	45.65	14	44.51	51.02
3	25.53	37.04	9	32.34	43.85	14	44.78	50.48
4	25.53	36.32	9	32.43	49.41	14	44.87	50.66
4	25.62	36.86	9	32.43	49.05	15	49.14	52.10
4	25.71	34.89	10	35.61	49.41	15	49.23	52.10
5	26.25	37.76	10	35.70	50.31	15	49.23	52.46
5	26.53	36.32	10	35.79	51.20	16	57.04	68.23
5	26.80	37.04	11	40.42	44.21	16	57.22	68.59
6	26.89	42.78	11	40.42	44.03	16	57.31	68.41

Table 2. Average Concentration Values and Their Corresponding Standard Deviations Obtained by Applying Both the Enzymatic (E) and the Somogyi–Nelson (S–N) Method (Values are Glucose in g L^{-1})

Sample #	Method E		Method S–N	
	Average	Standard dev.	Average	Standard dev.
1	13.72	0.09	22.52	0.95
2	23.89	0.09	36.20	3.01
3	25.34	0.18	37.70	0.85
4	25.62	0.09	36.03	1.02
5	26.52	0.27	37.04	0.72
6	27.04	0.19	42.78	1.08
7	31.25	0.16	42.96	1.61
8	31.79	0.09	45.29	0.62
9	32.40	0.05	47.44	3.11
10	35.70	0.09	50.31	0.90
11	40.45	0.05	44.09	0.10
12	41.03	0.14	48.57	0.27
13	43.96	0.07	49.95	0.18
14	44.72	0.19	50.72	0.27
15	49.20	0.05	52.22	0.21
16	57.19	0.14	68.41	0.18

Table 3. Statistical Parameters Obtained When Applying OLS, WLS, and BLS Statistical Techniques to Glucose Values Corresponding to 16 Samples Obtained by Two Analytical Methods

Technique	Statistical parameters			
	Intercept (a)	Confidence interval ($\alpha = 0.05$)	Slope (b)	Confidence interval ($\alpha = 0.05$)
OLS method	15.7	±6.5	0.84	±0.16
Conclusion	Presence of bias			
WLS method	– 0.3	±9	1.14	±0.20
Conclusion	Absence of bias			
BLS method	1.6	±9	1.09	±0.21
Conclusion	Absence of bias			

As can be observed in Figure 3, the BLS and WLS regressed lines significantly differ from the one corresponding to OLS. It is apparent that the data points with lower variance (corresponding to high glucose contents) are given comparatively larger weights in the least-squares fitting. This causes differences mainly in the calculated intercept.

Conventional individual confidence intervals for the slope and the intercept can lead to erroneous conclusions when carried out independently of each other because this ignores their strong mutual correlation. Instead of these individual tests, the elliptic joint confidence region (EJCR) for the slope and intercept is recommended [3, 6]. Figure 4 shows these regions calculated for OLS, WLS, and BLS techniques. The process is applied to know if the joint confidence interval test based on the regression technique provides correct results. New methodologies that show nonstatistical differences with respect to the reference method at the level of significance chosen must be accepted, and new methods that provide results that differ statistically from the results obtained using the reference method must be rejected [6]. It is important that for several cases the joint confidence interval test based on ordinary least-squares or weighted-least-squares tests provides results which differ significantly from the ones obtained with the joint confidence interval test based on the BLS. In the present work, the use of individual confidence intervals for the slope and intercept seems to indicate, both for WLS and BLS, that no bias exists, while for OLS it suggests the presence of bias (see Table 3); however, the consideration of the joint confidence regions drawn in Figure 4 allows one to reach a different conclusion: a bias is present for the three least-squares methods discussed.

Conclusions

The use of bioprocesses helps to present chemometrics to students as a real-life topic. A simple example is presented in this work, which generates the number of samples required for a proper comparison of two methods aimed at the determination of glucose. Implementation of this type of laboratory work of interdisciplinary character increases student's motivation and clearly improves the learning process.

Simple instruments and nontoxic reagents were used throughout this work, making it both safe and easy to apply in an undergraduate analytical chemistry course.

Acknowledgments. Financial support from the Universidad Nacional del Litoral (Projects CAI+D 34-2 and 219) and Fundación Antorchas is gratefully acknowledged.

Supporting Materials. Two supporting files, Matlab routines (supporting material 1.pdf) and experimental details (supporting material 2.pdf) are contained in one zip file (http://dx.doi.org/10.1007/ s00897020596b)

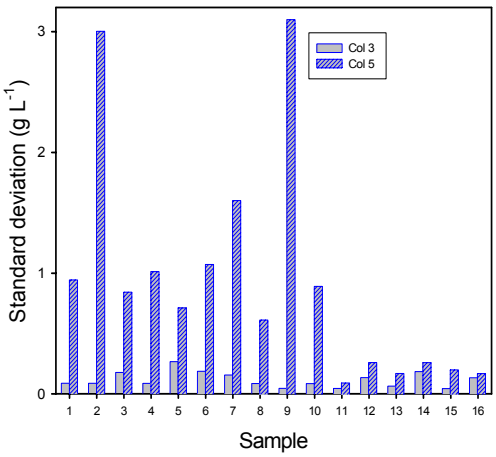

Figure 2. Standard deviation for both methods using different concentration levels.

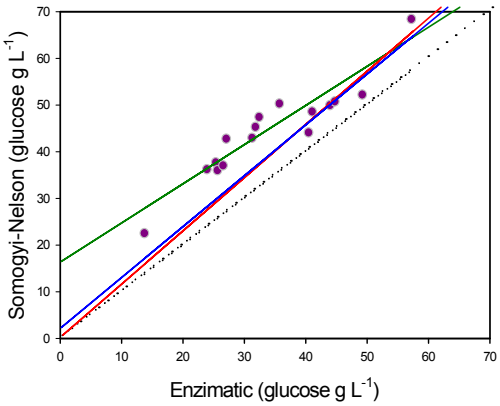

Figure 3. Concentration values when applying both analytical methods (spots), dotted line: ideal regression line corresponding to slope = 1 and intercept = 0, solid lines: OLS (green), WLS (red), and BLS (blue) regression lines.

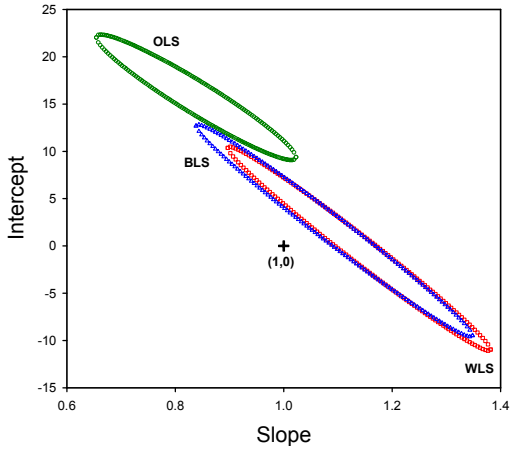

Figure 4. Joint confidence intervals based on OLS (green), WLS (red), and BLS (blue) methods ($\alpha = 0.05$).

References and Notes

1. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A.;* Elsevier, Amsterdam, 1997.
2. Valcárcel, M. *Principios de Química Analítica;* Springer-Verlag Ibérica: Barcelona, 1999.
3. González, A. G.; Herrador, M. A.; Suero, A. G. *Talanta* **1999,** *48*, 729–736.
4. Boqué, R.; Rius, F. X.; Massart, D. L. *J. Chem. Educ.* **1993,** *70,* 230–232.
5. Draper, N.; Smith, H. *Applied regression Analysis,* 2nd ed.; Wiley & Sons: New York, 1981.
6. Riu, J.; Rius, F. X. *Anal. Chem.* **1996,** *68,* 1851–1857.
7. Kansiz, M.; Gapes, J. R.; Mac Naughton, D.; Lendl, B.; Schuster, K. C. *Anal. Chim. Acta* **2001,** *438,* 175–186.
8. Frank, G. W. *Kombucha,* 1st ed.; Ennsthaler D. Steyr: city of publication, **1993,** pp 19–46.
9. Goicoechea, H.; Eluk, D.; Kubescha, M.; Ferraro, J.; Miglietta, H.; Rodil, B.; Mantovani, V. *Chem. Educator* [Online] **2000,** *5,* 67–70; DOI 10.1007/s00897990367a.
10. Lott, J.; Turner, K. *Clin. Chem*. **1975,** *21,* 174–1760.
11. Kaplan, L.; Pesce, A. *Clinical Chemistry Theory, Analysis and Correlation*, 1st ed.; The C. V. Mosby Company: St. Louis, Toronto, Princeton, 1986.
12. Holme, D.; Peck, H. *Analytical Biochemistry*, 2nd ed., Logman Scientific & Technical: England, 1993.
13. Lopez, P.; De la Ruente, J. L.; Borgos, J. *Anal. Biochem.* **1994,** *220,* 346–350