Available online at www.sciencedirect.com
ScienceDirect
Journal of Pharmaceutical and Biomedical Analysis 43 (2007) 1100–1105
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

Methodology of the thyroid gland disease decision-making using profiling in steroid hormone pathway

Young Sun Kim, Chang No Yoon *

Bioanalysis and Biotransformation Research Center, Korea Institute of Science and Technology, P.O. Box 131, Cheongryang, Seoul 130-650, Republic of Korea

Received 3 July 2006; received in revised form 18 September 2006; accepted 20 September 2006
Available online 1 November 2006

Abstract

To find out the genetic factors of outbreak of thyroid gland disease, we developed the thyroid gland decision-making system, which processes the metabolic profile in steroid hormone map using a statistical method. Metabolic profile is a measured data of lots of mixed materials that includes not only known metabolites, but also unknown ones, which is estimated to have an influence on the thyroid gland disease. Therefore, to develop thyroid gland disease decision-making system, analyzing metabolic profile containing multi-materials would be useful for diagnosing thyroid gland disease. Because experimental values used for system construction are area values for the retention time, the observations are preprocessed through variable transition and *t*-test to use the area values concurrently and the highly correlated materials are estimated by principal component analysis. The thyroid gland decision-making system developed through the logistic regression is an excellent system demonstrating 98.7% accuracy in the classification table.

Keywords: Thyroid gland disease; Steroid hormone pathway; Metabolite; Decision-making system; Principal component analysis; Logistic regression; 2-Hydroxyestrone; 2-Methoxyestrone; 2-Hydroxyestradiol

1. Introduction

To find out the genetic factors of outbreak of thyroid gland disease, we studied relationship between the disease and the thyroid gland disease decision-making system (TGDDM system) on the metabolic profile in the steroid hormone map.

The first approach on profiling was conducted to steroids in urine, sugars, sugar alcohols and organic acid participated in Krebs cycle [1] and the study on the steroid profile was carried out live [2–6] after introduction of analysis method using glass capillary column and GC–MS.

Furthermore, the constituent compound of the profiling was examined by the relative intensity and retention time information in LC–MS/MS [7], and data reduction is conducted through principal component analysis (PCA), which performs the pattern recognition in spectrum domain [8,9].

Contrary to the conventional researches, we analyze the metabolic profiling, which deals with the thyroid gland disease group and the normal group to develop the TGDDM system.

Conventionally, in the case of thyroid cancer, the effectiveness has been studied about the known metabolites, such as 2-hydroxyestrone, 2-hydroxyestradiol, 2-methoxyestrone, 2-methoxyestradiol, and 2-methoxyestradiol-3-methylether [10].

However, we examined a relation between the metabolic profiling and the thyroid gland disease, and developed the TGDDM system and its procedure.

We analyzed the recognition method of the spectral patterns using PCA as the disease diagnosis procedures using metabolic profile [7–9], we developed the diagnosis system, which determined the thyroid gland disease group and the normal group based on the PCA scores acquired from above analysis by estimating the thyroid gland disease discriminant model using logistic regression.

2. Materials and methods

Since the metabolic profile measured through LC–MS/MS is a spectrum type data, its variables must be translated for data analysis.

* Corresponding author. Tel.: +82 2 958 5068; fax: +82 2 958 5059.
E-mail address: cody@kist.re.kr (C.N. Yoon).

0731-7085/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2006.09.032

Retention time interval changes variously to estimate the discriminant mode of thyroid gland disease using intensity values measured at the retention time of the metabolic profile.

The variables are translated to estimate a proper statistic from the maximum, minimum, and mean values of the data measured simultaneously in each interval.

The translated variables are used for screening a test to determine whether they are risk factors of outbreak of thyroid gland disease, and the optimal variables are selected for the TGDDM system.

And also, the analysis dimension is reduced by principal component analysis extracting highly correlated variables as in existing papers [11].

Finally, we estimated the logistic model whose independent variables are PCA score and evaluate the model through the leave one out cross validation.

2.1. Data

Steroid and PUFA standards were purchased from Sigma (St. Louis, MO, USA). Isoprostane standards were obtained from Cayman Chemical Company (Ann Arbor, MI, USA). β-Glucuronidase/arylsulfatase from *Helix pomatia* was purchased from Boehringer Mannheim (Boehringer, Germany). As derivative reagents, *N*-methyl-*N*-trimethylsilyl-trifluoroacetamide (MSTFA), trimethylsilylchloride (TMCS), *N*-trimethylsilylimidazole (TMSI), pentafluorophenyldimethylsilyl (flophemesyl) chloride, and *N*,*O*-bis(trimethylsilyl)trifluoroacetamide containing 1% TMCS (BSTFA + 1% TMCS) were obtained from Sigma. Oasis HLB^{TM} cartridge was purchased from Waters (Milford, USA). All solvents were using a high-purified HPLC grade. De-ionized water was distilled before use.

We followed all of patients and normal samples at the Division of Endocrinology, Internal Medicine of Gyeongsang National University Hospital. We collected early morning urine samples from 10 normal controls, 14 hyperthyroidism patients, and 16 hypothyroidism patients. The mean age of female hyper- and hypothyroidism patients were 44.33 ± 17.37 (mean ± S.D.) years and 41.24 ± 15.60 years, respectively. Normal controls (mean age; 41.15 ± 16.41 years) were selected from the general population whom had no evidence of thyroid disease and taken a physical examination for this study. Hyperthyroidism was diagnosed on the basis of clinical features, diffuse goiter and positive antithyroid antibodies. The diagnosis of primary hypothyroidism was based on elevated serum TSH concentrations (basal, >4.1 mIU/l). We also found low serum concentrations of free thyroxine (free T_4), total T_4 (TT_4) and total triiodothyronine (TT_3). Using saturation analysis, we measured free T_4 (reference range, 12.0–28.0 pmol/l, Clinical Assays, Cambridge and NH). We measured the TT_4 (reference range, 60.2–160.0 nmol/l) and TT_3 (reference range, 1.2–3.1 nmol/l) by RIA as previously described (Clinical Assays). We measured the T_3 uptake using a routine method and we calculated the fT_4I by multiplying the T_4 concentration by the T_3 uptake result. The collected urine samples were stored at −20 °C until they were analyzed. We measured the creatinine concentration in the urine using the Jaffé method. Metabolic profile is a quantitative data acquired in test for the materials utilizing in metabolite urine [3]. The measured metabolite profile is consist of well-known hormones such as androsterone, dehydroepiandrosterone, DHEA, 2-hydroxyestradiol, 2-hydroxyestrone, 2-methoxyestrone and other unknown hormones.

Let the metabolic profile measured at time *j* on *i*th observation value X_{ij}.

Where the intensity of measured material denotes area and *j* denotes retention time.

Since the area value at retention time, which explains the characteristics of thyroid gland disease group and normal group is independent variable that is the most important in constructing the TGDDM system and it should be measured at the same reference time for all observations. The measured metabolic profile X_{ij} however, was not measured at the same reference time in each observation. Thus the same reference time is generated and given to each measured data X_{ij} for all observations. First, the measuring time point is changed to measureing time interval. Therefore, every observation value, *j* has area value at the measureing inverval $t = [t + \Delta,\ t - \Delta]$. Since there are several measuring time points in the measuring interval *t*, the data is translated by the following three statistic equations:

$$T_{it}^{\max} = \max[X_{il}, X_{il+1}, \ldots, X_{im}] \tag{1}$$

$$T_{it}^{\min} = \min[X_{il}, X_{il+1}, \ldots, X_{im}] \tag{2}$$

$$T_{it}^{\text{mean}} = \text{mean}[X_{il}, X_{il+1}, \ldots, X_{im}] \tag{3}$$

where $t - \Delta \le X_{il}, X_{il+1}, \ldots, X_{im} \le t + \Delta$.

The above calculated statistics denote the analytic variables $T_{it\,g}^{*}$ in specific interval of observation *j*.

Where g denotes the binary value representing the thyroid gland disease group and the normal group.

Therefore, incrementing the retention time interval Δ as 0.25, 0.5, 1.0, 1.5, 2.0, and so on, the interval value that is most suitable explains thyroid gland disease and the statistic should be estimated at the interval.

The screening test determines whether it is the risk factor to find out the optimal interval and its statistic.

2.2. Risk factor analysis

The *t*-test is executed to see whether the translated independent variable $T_{it\,g}^{*}$ is risk factor for outbreak of the thyroid gland disease. The *t*-test is a verifying method to see whether there is a difference between the two groups [12].

If $\bar{T}_0$, s_0^2 denote mean and variance of normal group of the thyroid gland disease, and $\bar{T}_1$, s_1^2 denote the treatment group, respectively, we can verify whether the specific interval in chromatography is risk factor of the outbreak of thyroid gland disease:

$$t = \frac{\bar{T}_0 - \bar{T}_1}{\sqrt{s_0^2/n_0 + s_1^2/n_1}} \tag{4}$$

The existence of difference between the thyroid gland disease group and the normal group is t-tested by applying various retention time and statistic. Where the analysis parameters are

mean, maximum, and minimum values at the translated interval of 0.5, 1.0, 1.5, and 2.0. By the result of the above *t*-test, we can estimate optimal interval value to differentiating the thyroid gland disease group and the normal group. And also, we can see among which statistic measure area values at the retention time interval is better to differentiate between the two groups.

2.3. Principal component analysis

Principal component analysis (PCA) is appropriate when you have obtained a measurement on a number of observed variables and wish to develop a smaller number of hypothetical variable (called principal components) that will account for most of the variance in the observed variables. The principal components may then be used as predictor or criterion variables in subsequent analyses.

The hypothetical variable is also called as a latent variable, a theoretical variable and a construct variable.

In this research, investigation of the Pearson correlation on the observed variables, which is constructed based on the retention time interval shows that lots of variables seem to be highly correlated to each other.

Thus we can calculate the principal component score through principal component analysis on the observed variable T^*_{ij}, which is generated based on retention time interval.

In some previous researches [8,9], they made efforts to clarify the type of materials by estimating the retention time interval of high correlation using PCA.

2.4. Discriminant model

Discriminant model is a model to make an estimation and a prediction about the causal relation based on the independent variables, in case the response variable is a categorical type.

This method, however, can not use the logistic regression in case the response variable is a binary type as the existence of disease.

If y denote variable to represent thyroid gland disease or not (1: disease; 0: normal), and denote $T^*_{it\,\mathrm{g}}$ is statistic at the translated retention time interval, the logistic regression model can be expressed as Eq. (5).

Estimation of parameter coefficient β_i can complete the model:

$$\log \frac{p(y=1|T^*_{it,\mathrm{g}})}{1-p(y=1|T^*_{it,\mathrm{g}})} = \alpha + \beta_i \times T^*_{it,\mathrm{g}} \tag{5}$$

Since the logistic regression is linear and has parametric characteristic, the performance of prediction on the new data maintains to some extent. On the contrary, the data mining method about the non-linear and non-parametric models has been studied profoundly. The representative examples are neural network, decision tree, support vector machine and so forth [13]. This model has high prediction performance, but it has shortcoming to be over-fitted to training set of model estimation. Comparative evaluation should be performed to get an optimal model and it should be verified that the selected model is superior to other ones. In this paper, we estimate a model optimal onto thyroid gland disease prediction system using above various models. As a result, validation sets are estimated 100% for all models. Therefore, the thyroid gland disease dicriminant model is estimated using logistic regression model that is easily analyzable and applicable.

2.5. Model validation

The first criterion of model validation is how few independent variables are used to construct effective model. In this paper, we utilize screening analysis to select optimized variables by *t*-testing the explanatory variables and reduce the variables by PCA. The second criterion of model validation is how a stable result is produced when the model is applied to a new data. In other words, it means the possibility of generalization of the model. However excellent the prediction performance of the model is, it is useless if it can not be generalized. In this paper, since the number of data was not enough to be partitioned into the training set and the validation set, we verified using leave one out cross validation [14]. Leave one out cross validation is a method to utilize one of n data set as validation set and the other $n-1$ data sets as training set, so we can validate all the n data sets through n models. It can be used to analyze few data sets as in this case.

3. Result

3.1. Translation variable and risk factor

The variables are generated by varing retention time's interval. Although several area values are included in the generated variables, only one statistic should be selected and used. Thus we need to decide the criterion to select the statistic and the interval size. The *t*-test is used to select the statistic and the interval size that explain the characteristics of the treatment group and the normal group. We could get the results of Figs. 1–3 by changing the retention time interval from 0.5 to 3.0 variously and applying corresponding statistic to the *t*-test.

In the figures, *x*-axis denotes the measured retention time, *y*-axis denotes translated interval, and *z*-axis denotes *p*-value. Therefore, the lower the values of *z*-axis, the better the variables explain characteristics of the treatment group and the normal group. In Fig. 1, the result using mean value at each interval as a statistic is shown, in Fig. 2, maximum value, and in Fig. 3, minimum value is used. As the findings of figures, there exist area values that can effectively discriminate between the treatment group and the normal group over 18 of retention time irrelevant to statistic.

Also**,** overall distribution patterns show that the least p-value is at the interval 1.5.

Therefore, the estimate of the optimal retention time interval becomes 1.5, and as a statistic, the maximum area value in that interval can be an important independent variable that explains characteristic of the outbreak of the thyroid gland disease.

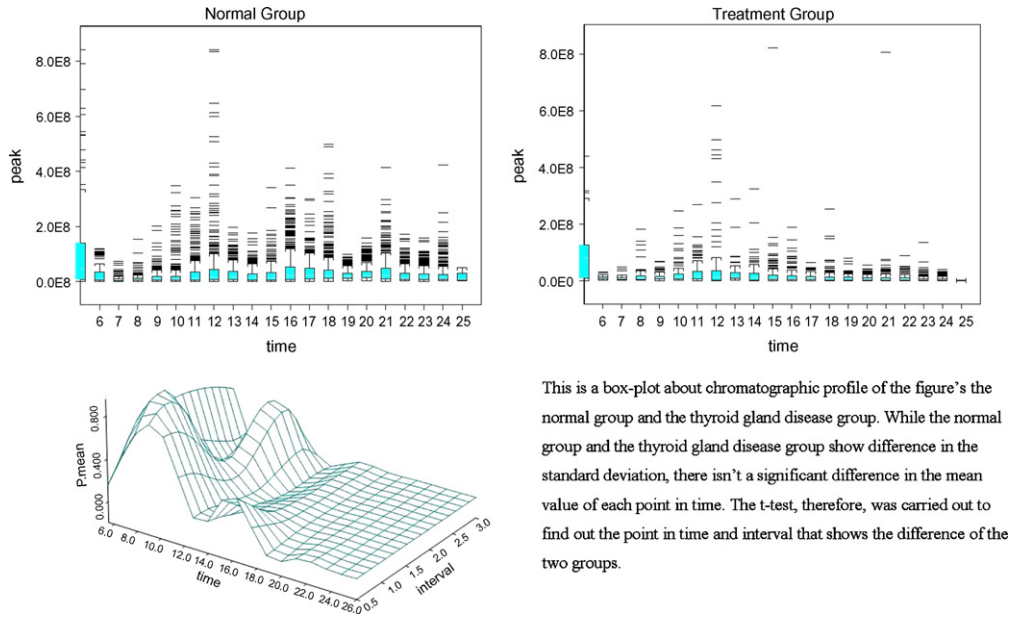

Fig. 1. Distribution of *p*-value with repect to mean value.

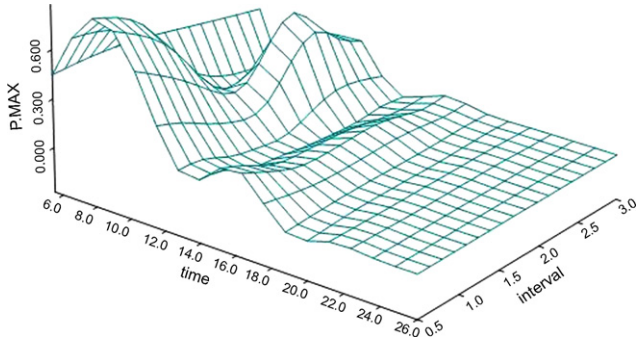

Fig. 2. Distribution of *p*-value with repect to maximum value.

3.2. Principal component analysis

A good analysis uses small amount of variables to explain their maximum variance. Thus in order to reduce the number of highly correlated independent variables, principal component analysis is performed. The number of principal component is determined by choosing eigenvalues that is greater than 1, or determined with respect to the cumulative variance.

Based on this consideration, the number of principal component is determined to be 3, and by this number of principal component, around 89.6% of the data variance is explained as shown in Fig. 4.

It is shown that the first principal component has a high correlation with the variables of peak 13.5, peak 19.5, and peak 16.5, which is in the retention interval between 12 and 13.5 (Fig. 5).

The second principal component has negative correlation with peak 13.5 and has relatively higher correlation with peak 22.5 and peak 16.5. We can see that the third principal component has positive correlation with peak 22.5 and has negative correlation with 16.5.

The three principal components are explaining 90% of total variance of 13 explanatory variables.

Thus we showed how the three principal components are explaining the thyroid gland disease.

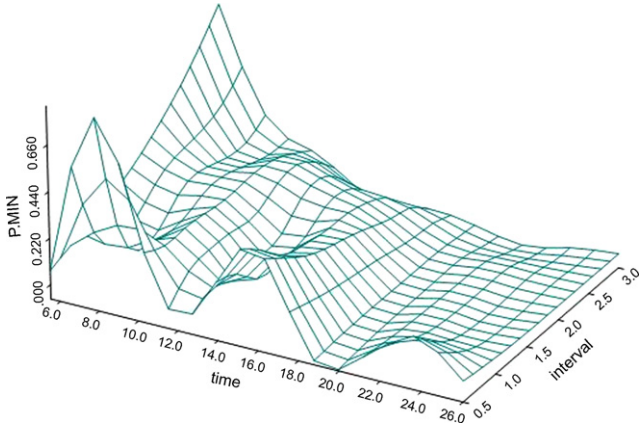

Fig. 3. Distribution of *p*-value with repect to minimum value.

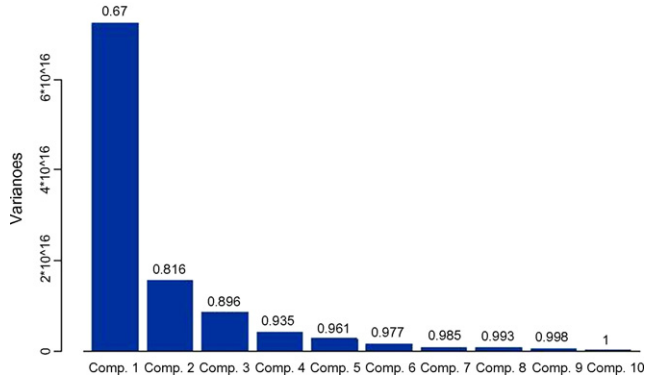

Fig. 4. Relative importance of principal components.

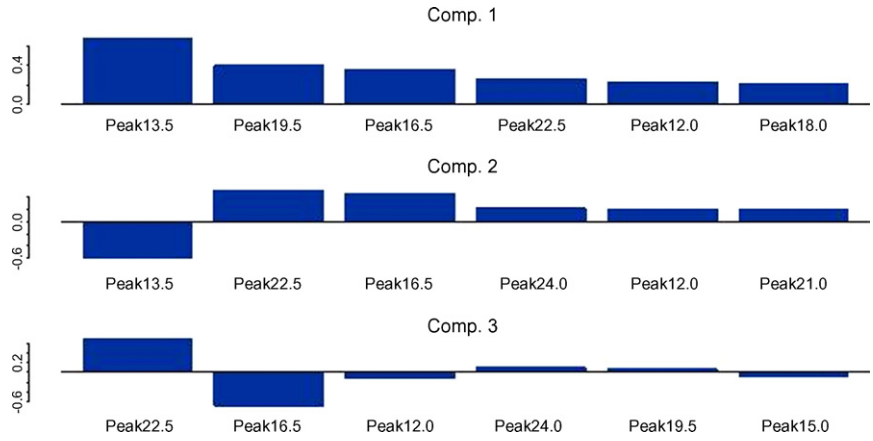

Fig. 5. Principal component score.

To see the distribution of disease group and normal group about the reduced independent variables of three principal components, it is plotted in a three-dimensional space as shown in the Fig. 6.

First of all, the disease group and normal group are distinguishably distributed in a three-dimensional space.

As a result, the normal group ($\bar{C}1_0 = 1.19$) is distributed more than the disease group ($\bar{C}1_1 = -1.37$) depending on how large the first principal component is.

For second principal component, on the contrary, the larger the value is, the more treatment group ($\bar{C}2_1 = 0.30$) is distributed than the normal group ($\bar{C}2_0 = -0.26$). And for third principal component, the larger the value is, the more normal group ($\bar{C}3_0 = 0.24$) is distributed than the treatment group ($\bar{C}3_1 = -0.28$).

Thus, from the measured value area of the variables generated at retention time interval 1.5, we can conclude that there are many unknown metabolites in the metabolic profile**,** which take a similar effect on the thyroid gland disease and it can be estimated that these materials are significant to the outbreak of the thyroid gland disease.

3.3. Logistic regression and validation

We validated discriminant models of logistic regression, neural network, and decision tree, which discriminate between normal group and treatment group based on the principal component score of above analysis.

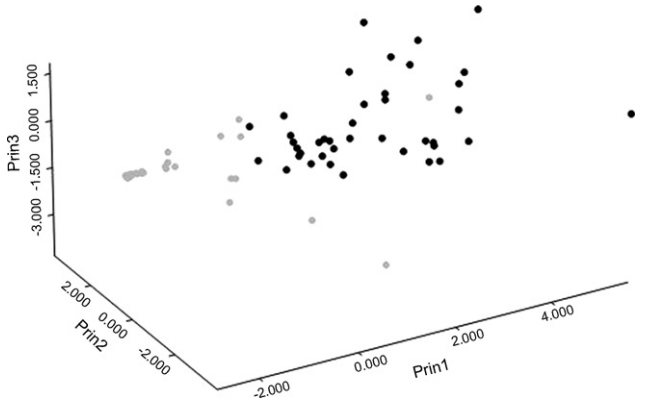

Fig. 6. Distribution of the disease group and the normal group with the axes of three principal components.

Table 1
Result of logistic regression

Parameter	Normal	Treatment	Estimate	*p*-Value
Intercept	–	–	−0.4866	0.3666
Prin				
1	1.19 ± 2.17	−1.37 ± 1.18	−1.7002	<.0001
2	−0.26 ± 1.68	0.30 ± 1.91	0.8672	0.1613
3	0.24 ± 1.51	−0.28 ± 1.36	−1.3104	0.0089

Since all three models have the accuracy of 98.7% in validation set of classification table, the logistic regression model among 3, which is easy to analyze, is used to construct thyroid gland disease discriminant system (Table 1).

From the analysis of logistic regression with input of three principal components that are explaining 90% of variance of explanatory variables, it showed that the first principal component had a parameter coefficient of −1.70 and took the largest effect on the model.

Then the coefficient of the third principal component was −1.34 and the second principal component was 0.86.

The model accuracy from leave-one-out cross validation revealed that the observation of one treatment group was missclassified, and other observations was classified properly. Thus we developed the TGDDM system using logistic regression with input of measured three principal component.

4. Discussion

Conventionally, the studies on outbreak cause of thyroid gland disease were concentrated to the environmental effects. However, though people had the same environmental position, there was a significant difference in the outbreak of the disease. This was the reason why this phenomenon could be caused by genetic factor.

More researches are required to find out which hormone influences the outbreak of the disease.

There is not any clarified or completed study result on metabolites in steroid map.

Therefore, we have studied the TGDDM system using the metabolic profile, which includes measured value of unknown metabolites.

The parallel research on the known metabolites affecting thyroid gland disease reveals that the hormones such as 2-hydroxyestrone, 2-methoxyestrone, 2-hydroxyestradiol, 2-methoxyestradiol, and 2-methoxyestradiol-3-methylether are the important risk factors of outbreak of the disease, but it merely means they are just a part of the risk factors, not all the factors. This is the reason why we are developing the TGDDM system considering all the materials through metabolic profile.

With respect to retention time interval size, the number of unknown materials in the interval is different. In this research, we classified the data by translating variables and t-test and constructed the TGDDM system using the logistic regression, which takes above classified variables as a explanatory variable in order to properly find out materials that explain the characteristics of the thyroid gland disease.

In the further study, we are going to emphasize our research on clarifying unknown materials that are included in the system.

References

[1] E.C. Horning, M.C. Horning, A. Zlatkis (Eds.), Advances in Chromatography, University of Houston Press, Houston, 1970, p. 226.
[2] C.H.L. Shachleton, J.A. Gustafsson, F.L. Mitchell, Acta Endocrinol. (Copenh.) 74 (1973) 157–167.
[3] J.A. Luyten, G.A. Rutten, J. Chromatogr. 91 (1974) 393–406.
[4] H. Ludwig, J. Reiner, G. Spiteller, Chem. Berichte. 110 (1977) 217.
[5] C.D. Pfaffenberger, E.C. Horning, J. Chromatogr. 112 (1975) 581–594.
[6] C.D. Pfaffenberger, E.C. Horning, Anal. Biochem. 80 (1977) 329–343.
[7] H. Yamanaka, M. Nakajima, K. Nishimura, R. Yoshida, T. Fukami, M. Katoh, M. Yokoi, EUFEPS 22 (2004) 419–425.
[8] H. Wu, X. Zhang, X. Li, Z. Li, Y. Wu, F. Pei, Anal. Biochem. 340 (2005) 99–105.
[9] A.M. Laverman, M. Braster, W.F.M. Röling, H.W. Verseveld, Soil Biol. Biochem. 37 (2005) 1581–1588.
[10] K.M. Kim, B.H. Jung, D.S. Lho, W.Y. Chung, K.J. Paeng, B.C. Chung, Cancer Lett. 202 (2003) 173–179.
[11] C.H.L. Shacketon, J. Chromatogr. 379 (1986) 91–156.
[12] B. Rosner, Fundamentals of Biostatistics, 4th ed., Duxbury Press, California, 1995.
[13] Y.S. Kim, S.Y. Sohn, C.N. Yoon, Biomed. Pharmacother. 57 (2003) 482–488.
[14] G.C. Cawley, N.L.C. Talbot, Pattern Recognit. 36 (2003) 2585–2592.