Drug Development and Industrial Pharmacy, 26(2), 123–134 (2000)

RESEARCH PAPER

Ruggedness Study of HPLC Peptide Mapping for the Identity of a Drug Compound: A Chemometrics Approach

Kwan R. Lee,[1,*] Jacob Bongers,[2] Brian H. Jones,[2] and Sudhir Burman[2]

[1]Statistical Sciences Department and [2]Bioanalytical Sciences Department, SmithKline Beecham Pharmaceuticals, 709 Swedeland Road, King of Prussia, PA 19406

ABSTRACT

A statistically more reliable approach than the traditional visual inspection of peptide maps to identify a drug compound is to generate a set of reference standards from a designed experiment that incorporates many possible factors that affect variation of peptide mapping. In fact, the experiment can be done for a ruggedness study as part of a high-performance liquid chromatography (HPLC) method validation. Once the ruggedness is proved with the study, those articles in the experiment may form a set of reference standards, and future articles can be compared to the set later to prove identity. A quantitative analysis of the ruggedness study can be done using a chemometrics approach, principal component analysis (PCA). The analysis is used to reduce the many channels of peptide maps to a few manageable dimensions. The scores projected onto the reduced dimensions are used to test factor effects of the ruggedness study. As a by-product, the analysis provides visual inspection of the set of articles in the experiment for any outliers and anomalies.

Key Words: *Peptide mapping; Principal component analysis; Ruggedness study.*

* To whom correspondence should be addressed.

INTRODUCTION

Peptide mapping is a powerful technique for studying the primary structure of proteins. The method generally employs site-specific proteases to generate unique peptide fragments, which are then analyzed by a separation method, such as reversed-phase high-performance liquid chromatography (RP-HPLC). The sensitivity of the peptide map to even the smallest change in the covalent structure of the protein makes it a valuable ''fingerprint'' for identity testing and process monitoring. For recombinant protein pharmaceuticals, peptide mapping is used for the initial ''proof-of-structure'' characterization, that is, to confirm expression of the desired amino acid sequence and to characterize any posttranslational modifications such as glycosylation, proteolytic processing, acetylation, and so on. Further, peptide mapping is employed for subsequent lot-to-lot identity testing (fingerprinting) in support of bioprocess development and clinical trials. Peptide mapping is also the current method of choice for monitoring ''genetic stability.'' Finally, the method is validated to meet stringent federal regulations and transferred to a manufacturing site for quality control.

There are several recent reviews (1–5) on methods and methods validation for peptide mapping as applied to well-characterized biopharmaceuticals. We recently conducted a full method validation study of an optimized RP-HPLC tryptic map of a therapeutic immunoglobulin G1 (IgG1) monoclonal antibody (6). We have used this method routinely for over a year to test production lots for ongoing clinical trials and to support bioprocess development. For certain proteins, including antibodies, it is necessary to remove intrachain disulfide linkages first in order for trypsin to fragment or ''digest'' the protein. Trypsin cannot fragment the protein substrate unless these covalent disulfide bridges are broken. This is generally accomplished by a reduction/alkylation of the cysteine residues in the protein. Thus, our method consists of an initial reduction with the dithiothreitol (DTT) reagent, alkylation with iodoacetate reagent, trypsin digestion, and, finally, an RP-HPLC analysis of the peptide fragments. These methods were examined in the ruggedness study; the method was varied to different degrees to determine these effects on the peptide map of the antibody.

To inspect each chromatogram, it was necessary to convert the pictorial representation that the chromatographic software (HP ChemStation, Hewlett-Packard, Waldbronn, Germany) output into a usable numeric form. We were able to obtain these desired data utilizing a macro that was written by Hewlett Packard (R. Giuffre). This macro, known as DATAOUT9, sends retention times and the corresponding response data to an ASCII file. However, each chromatogram is made up of approximately 30,000 individual retention times and corresponding response values. To inspect the chromatographic data quantitatively, it was necessary to take these numerous data points and reduce them to a more usable form. Through some programming efforts of our own, we were able to adapt the macro to take every *k* consecutive data points, average the responses, and then output the average value to an ASCII file. These response values were then used to give a numeric representation of each chromatogram that was unique enough to represent the variation that existed in the pictorial representation of the chromatograms.

To identify a drug compound, often a visual inspection of an HPLC peptide map is made by comparing it to the reference standard. Sometimes, a test article is put side by side with the reference standards, and every peak is examined (Fig. 1). This approach ignores random varia-

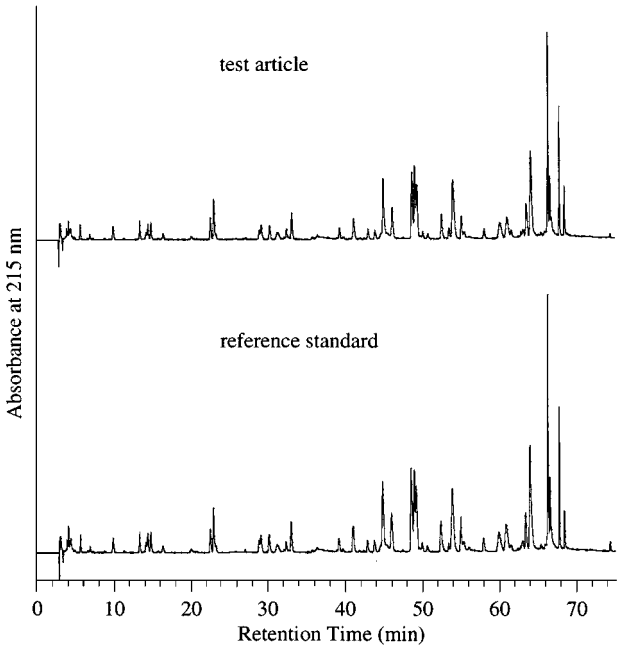

Figure 1. Typical reversed-phase high-performance liquid chromatography (RP-HPLC) tryptic mapping data for a monoclonal antibody protein. The proteolytic enzyme trypsin is used to ''cleave'' the protein molecule into a set of specific peptide fragments (i.e., tryptic fragments). The RP-HPLC analysis then provides a profile or ''fingerprint'' that is characteristic of the particular protein amino acid sequence (primary structure) and exquisitely sensitive to any small change in chemical structure. Identity testing involves comparison of the map of the test article to that of a fully characterized and carefully preserved reference standard prepared in parallel.

tion entered in the peptide maps and, at best, is not very repeatable since it can rarely answer the question of how different is different. A statistically more reliable approach is to generate a set of reference standards from a designed experiment that incorporates many possible factors that affect the variation of peptide maps. Better yet, a ruggedness study can be done based on a designed experiment as part of a method validation of HPLC peptide mapping. Once the ruggedness is proved with the study (i.e., none of the factors in the experiment significantly affects the variation of the maps), those articles included in the experiment can be declared reference standards, and future articles can be compared to the set of standards to prove their identities.

One of the technical difficulties in this ruggedness study is the fact that the response of the experiment is the peptide map itself; it has many channels of information (typically more than 10,000 channels). One idea is to reduce the peptide map to several summary measures, which may include several chosen peaks and areas under the peaks. Again, this approach is still time consuming and operator sensitive, and the method is rarely transferable. A chemometrics approach using principal component analysis (PCA) (see, e.g., 7) may effectively reduce many channels of data into manageable dimensions and as a by-product provide visual inspection of all the articles in the experiment in a simpler, reduced space. The inspection in the reduced dimension may be used to pinpoint an outlier and other anomalies of the experiment.

The projected scores on the reduced dimensions can be effectively used as summary measures to complete the ruggedness study. The goal of this paper is to illustrate the approach step by step using an actual example in drug development.

EXPERIMENTAL

The IgG1 monoclonal antibody used for this study was provided by SmithKline Beecham.

Reduction/alkylations were done by evaporating the antibody sample to dryness, dissolving the residue in 6 M guanidine hydrochloride, 1.2 M Tris/HCl buffer (pH 8.1) to 10 mg/ml, reducing with 50 mM DTT at 65°C for 30 min, alkylating with 120 mM sodium iodoacetate at room temperature for 40 min in the dark, and then desalting the carboxymethylated antibody into 50 mM Tris/HCl, 1 mM $CaCl_2$ (pH 8.1) digestion buffer by use of a prepacked disposable gel-filtration column (Bio-Rad Econo-Pac 10DG Bio-Gel P-6 column).

Tryptic digestions of the resulting 2.0 mg/ml carboxymethylated antibody in 50 mM Tris/HCl, 1 mM $CaCl_2$ (pH 8.1) were done with 0.02 mg/ml trypsin (Worthington Biochemical Corp., Lakewood, NJ, bovine pancreatic, treated with 1-tosylamide, 2-phenylethyl, chloromethyl ketone [TPCK]) at 37°C for 2 hr. Digestions were halted by acidifying to approximately pH 2 with 1.0 M HCl.

RP-HPLC was performed on a Hewlett Packard 1090M LC with a diode-array ultraviolet (UV) detector using the following: Vydac 218TP54 C18 column (4.6×250 mm) thermostated at 35°C; the eluents (A) 0.1% TFA and (B) 0.1%TFA in 80% v/v CH_3CN; the gradients 5% to 40% B in 60 min and 60% to 100% B in 20 min at 1.0 ml/min; and detection at 215 and 280 nm.

PRINCIPAL COMPONENT ANALYSIS OF PEPTIDE MAPS

Original data used in this example had more than 30,000 channels in each peptide map. To make the data analysis more manageable, data were reduced by picking 1 from every 10 channels of information. In this way, data were reduced effectively, but contained enough details of the original data. We have also used averages (e.g., averages of five consecutive channels) and observed very similar results. An approach most commonly used is to transform the point representation of the profiles to peak heights or peak areas, but this is time consuming, operator intensive, and often is inaccurate for peak identification and resolution.

Figure 2 shows the average of 18 maps. It has 3188 channels (variables). Using PCA, we drastically reduce the dimensions to a few manageable numbers using PCA.

Subtracting the average map from each of the 18 peptide maps (centered) was the initial pretreatment of the data. The 18 maps were generated from a designed experiment, which is explained in the ruggedness study section below. Figure 3 shows the centered case 5-1, which was somewhat atypical among 18 cases since most of the channels had negative readings. Typically, data are centered before multivariate analysis of spectroscopic or chromatographic data since we are mostly interested in the variation from the averages. However, scaling of the data is rarely used for same type of data since all channels already have comparable scales.

The chemometrics software SIMCA® (8) was used to perform PCA, and Table 1 shows the essential summary statistics from the model building. The first two components explain more than 90% of the variation (R^2, cumulative) and more than 86% of the cross-validated variation in the data (Q^2). Cross validation means that, when

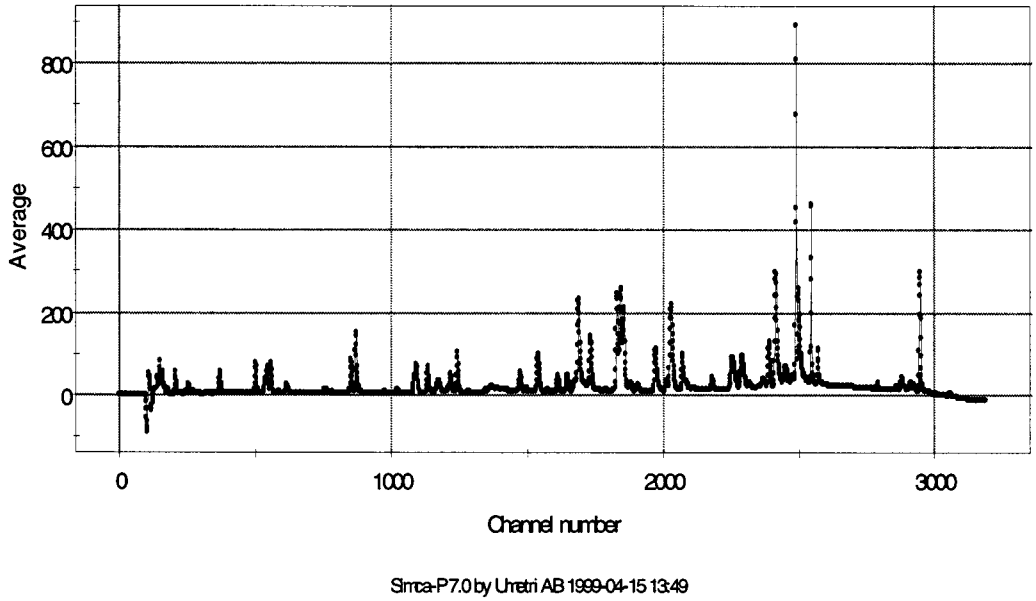

Figure 2. An averaged HPLC peptide map of monoclonal antibody.

each case was predicted, the particular case was not used in the model building; hence, Q^2 can be considered to be a more realistic measure of goodness of fit than R^2. Although all seven components may be used for future prediction purposes (they were statistically significant), the first two dimensions may be adequate for exploratory analysis of the data and its visualization. The eigenvalue can be interpreted as a measure of overall variation in each dimension. The larger the eigenvalue is, the more information is in the dimension.

Now, we are ready to examine the reduced two-dimensional space for overall scatter of the data and some

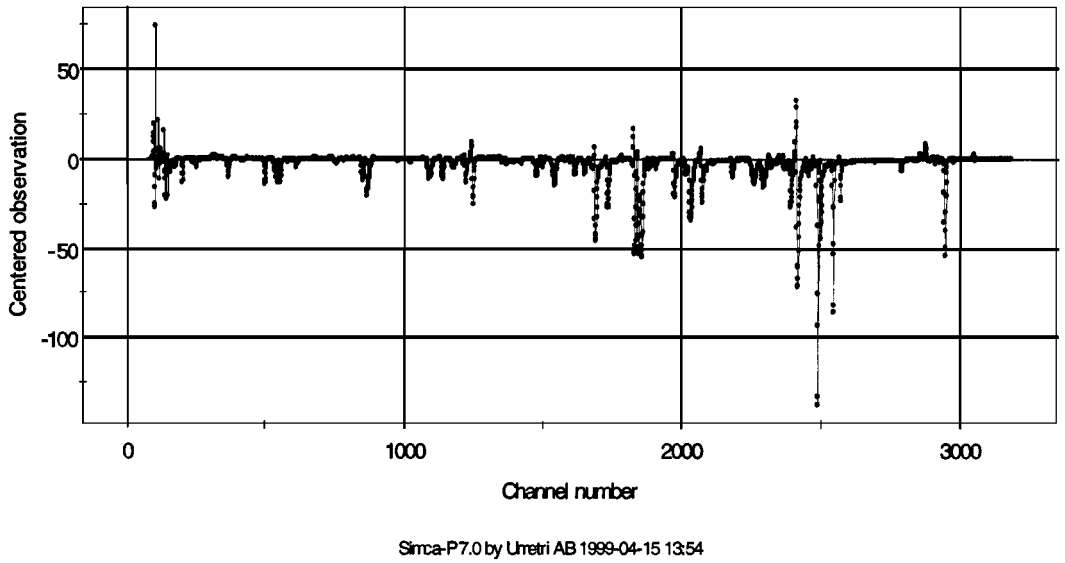

Figure 3. A centered peptide map of an unusual case.

Table 1

Summary Statistics from the Principal Components Analysis

A	R^2	R^2 (cumulative)	Eigenvalue	Q^2	Q^2 (cumulative)
1	0.843	0.843	15.173	0.818	0.818
2	0.063	0.906	1.139	0.253	0.864
3	0.025	0.932	0.456	0.095	0.877
4	0.020	0.952	0.364	0.130	0.893
5	0.016	0.968	0.294	0.239	0.919
6	0.008	0.976	0.139	0.119	0.928
7	0.006	0.982	0.111	0.118	0.937

possible anomalies. Figure 4 shows the projected scores of the first two dimensions. Note that the second numbers in the label designate replicates. For example, 3_1 and 3_2 are replicates of experiment 3. Most variation comes from replicate-to-replicate variation, and experiment-to-experiment variations are smaller. Also, case 5_1 is well outside the 95% confidence ellipsoid determined by Hotelling's T^2. There seems to be no appreciable systematic variation due to factor effects. This is examined in more detail in the ruggedness study section.

Lower dimensional projection of the data as in Fig. 4 may be most useful for identifying outliers and possible groupings of the data (cluster analysis). Another useful companion to the scores plot is the loadings plot. Figure 5 shows the two-dimensional loadings plot based on all 18 peptide maps. If a particular variable is far away from the origin of the plot, then we might say that particular variable is more important than the others in computing the reduced scores. The loadings can be interpreted as weights to the original variable or directional cosine of the original variables. For example, V2487 to V2490 have distinct loadings away from others; hence, they are important channels for studying the variation of peptide maps. Sometimes, the loadings plot can be used together

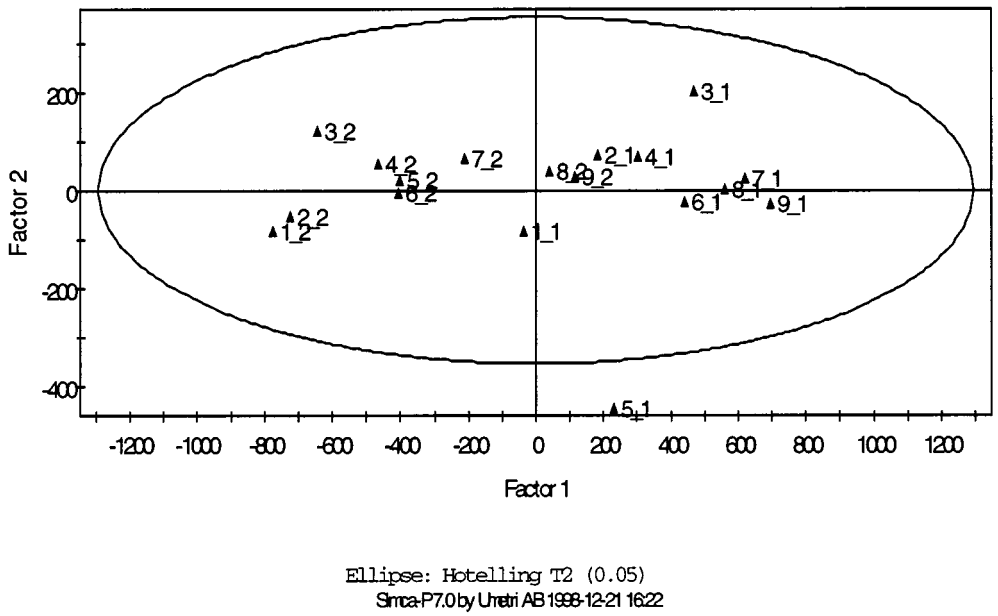

Figure 4. Projection of data onto the first two principal axes.

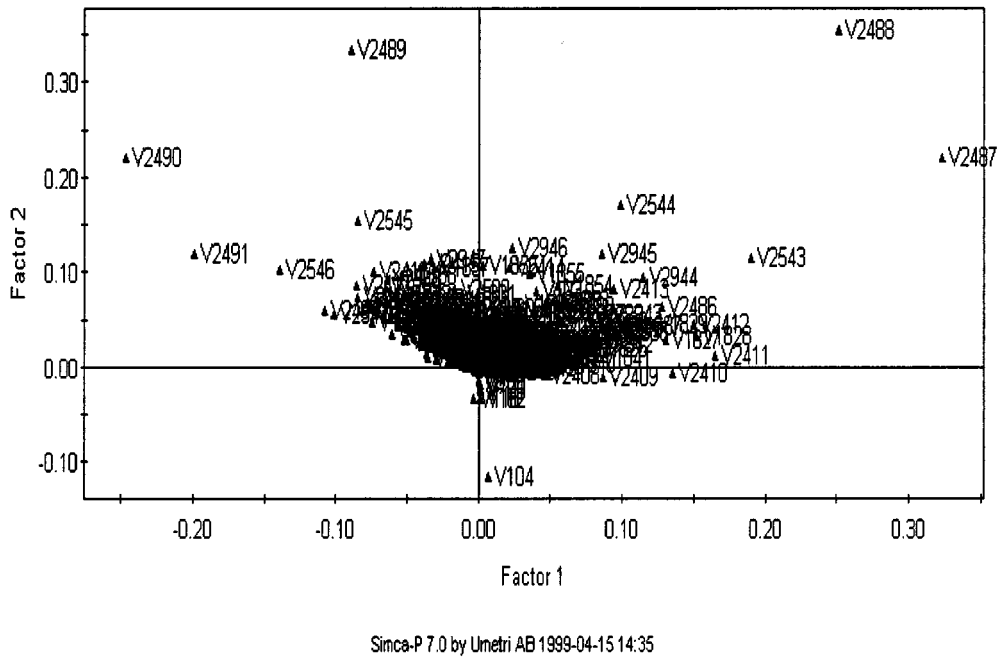

Figure 5. Loadings plot of the first two dimensions.

with the scores plot using the same origins (called a biplot) to know why some cases are different from others. For example, case 3_1 is most extreme in the first quadrant, and the loadings plot indicates it should have larger values in V2488 and in V2487. Also, the loadings plot suggests that case 5_1 is an outlier since most of its channels have negative numbers. This can be verified from Fig. 3.

The first two principal component scores are given Table 2. These numbers are used in place of the original data for the ruggedness study in the next section.

Table 2

Projected Scores of the First Two Principal Components

ObsNum	ObsName	Score1	Score2
1	1_1	−35.09	−80.86
2	1_2	−775.11	−81.70
3	2_1	183.69	73.44
4	2_2	−721.82	−53.50
5	3_1	471.01	204.03
6	3_2	−643.85	123.57
7	4_1	301.79	68.22
8	4_2	−464.50	57.01
9	5_1	230.75	−445.25
10	5_2	−400.36	20.85
11	6_1	440.76	−21.62
12	6_2	−407.36	−2.66
13	7_1	619.28	27.25
14	7_2	−209.38	66.16
15	8_1	560.18	2.59
16	8_2	37.90	38.66
17	9_1	695.74	−27.22
18	9_2	116.35	31.00

RUGGEDNESS STUDY OF THE HPLC PEPTIDE MAPPING

So far, we have talked about principal component analysis as an essential tool to reduce higher dimensional data and to visualize the results. However, the main purpose of the dimensional reduction was to use the resultant scores for the ruggedness study of the HPLC method of peptide mapping. A fractional-factorial ruggedness study was designed to investigate the effects on the RP-HPLC mapping results of varying the amounts of the three reagents (DTT, iodoacetate, and trypsin), the three times of reaction, and finally, the temperature of the enzymatic

Table 3

Factors and Their Settings in the Ruggedness Study

Factor Label	Meaning	Unit	Low	High
A: DTT	Dithiothreitol solution	µmol	22.50	27.50
B:Iodoacetate	Sodium iodoacetate solution	µmol	54	66
C: Reduction *t*	Time of reduction	min	20	40
D: Alkylation *t*	Time of alkylation	min	30	50
E: S/E	Substrate/enzyme		90	110
F: Digestion *T*	Digestion temperature	°C	35	39
G: Digestion *t*	Digestion time	min	100	140

digestion step. The seven factors included in the ruggedness study are shown in Table 3.

To estimate the main effects of all seven factors properly, a fraction (1/16) of two-level full-factorial design (see, e.g., Ref. 9) was used with replicates. Also, two center runs at standard conditions were added to the 16 runs. Those 18 runs were randomized in the actual experiment. Table 4 shows the full design with the scores.

We analyzed the data using Design-Expert® (10), a statistical software package developed especially for designing and analysis of experiments. The main effects of all seven factors were calculated, but they did not significantly affect the scores, as is shown in the half-normal plots (Figs. 6 and 7) and *t* tests of the effects (Tables 5 and 6).

The triangles in Figs. 6 and 7 indicate pure error effects from replicates, and the labeled solid squares are the effects of the factors. If the effects of factors are considerably larger than pure error effects, then we may see some significant effects (i.e., the solid squares would lie to the right of the line). In both figures, however, the effects were comparable to, or even smaller than, pure errors.

Hence, we may conclude the method is rugged within

Table 4

Designed Experiment and Two Principal Components Analysis Scores as Response

Standard	Run	A	B	C	D	E	F	G	Score1	Score2
1	3	22.5	54	20	50	110	39	100	695.74	−27.23
2	5	22.5	54	20	50	110	39	100	116.35	31.00
3	11	27.5	54	20	30	90	39	140	440.76	−21.62
4	14	27.5	54	20	30	90	39	140	−407.36	−2.66
5	6	22.5	66	20	30	110	35	140	471.01	204.03
6	12	22.5	66	20	30	110	35	140	−643.85	123.57
7	18	27.5	66	20	50	90	35	100	301.79	68.22
8	2	27.5	66	20	50	90	35	100	−464.50	57.02
9	17	22.5	54	40	50	90	35	140	230.75	−445.25
10	16	22.5	54	40	50	90	35	140	−400.36	20.85
11	10	27.5	54	40	30	110	35	100	619.28	27.26
12	8	27.5	54	40	30	110	35	100	−209.38	66.16
13	4	22.5	66	40	30	90	39	100	560.18	2.59
14	15	22.5	66	40	30	90	39	100	37.91	38.67
15	1	27.5	66	40	50	110	39	140	−35.10	−80.86
16	7	27.5	66	40	50	110	39	140	−775.11	−81.70
17	9	25.0	60	30	40	100	37	120	183.69	73.45
18	13	25.0	60	30	40	100	37	120	−721.82	−53.50

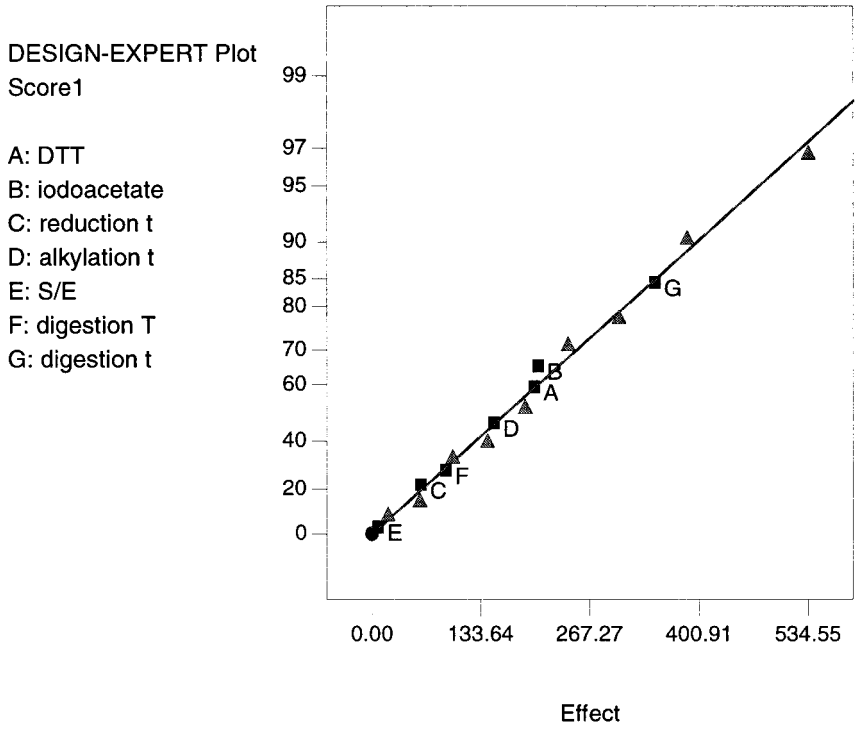

Figure 6. Half-normal plot of main effects for Score1.

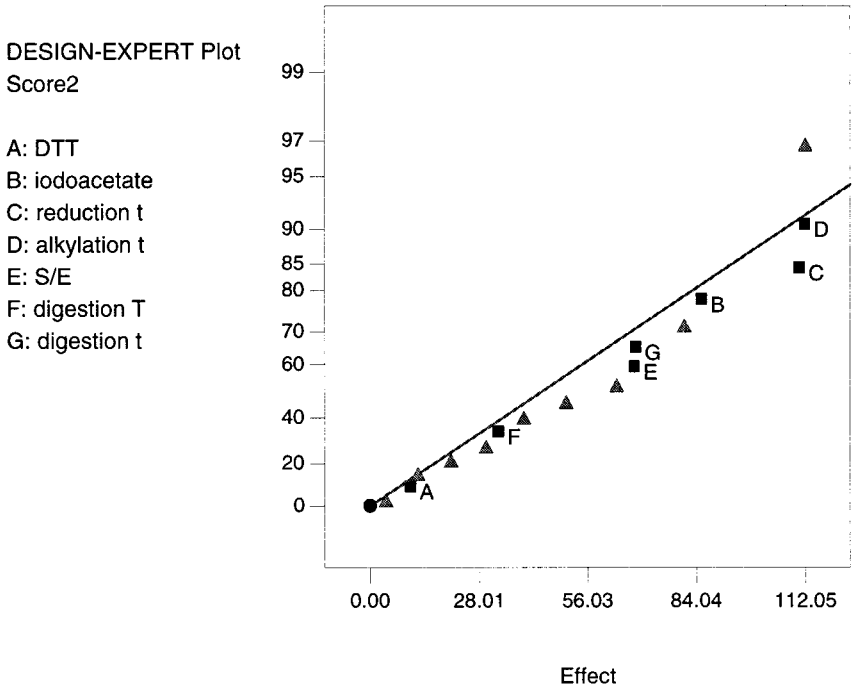

Figure 7. Half-normal plot of main effects for Score2.

Table 5

Effects Table for Score1

	Coefficient Estimate	Standard Error	t for H_0 Coefficient = 0	Prob > |t|
Intercept	33.63	139.57		
A DTT	−99.83	139.57	−0.72	0.4926
B Iodoacetate	−102.09	139.57	−0.73	0.4831
C Reduction *t*	−30.11	139.57	−0.22	0.8340
D Alkylation *t*	−74.94	139.57	−0.54	0.6044
E S/E	−3.76	139.57	−0.027	0.9791
F Digestion *T*	45.54	139.57	0.33	0.7517
G Digestion *t*	−173.54	139.57	−1.24	0.2451

the boundaries of the experiment and consequently use the 18 runs used in the experiment as a set of reference standards in the future.

IDENTITY OF PEPTIDE MAPS: USING REFERENCE STANDARDS

We obtained a set of reference standards from the ruggedness study discussed above; as mentioned, this set can be used in the future to prove the identity of the same drug compound later. To illustrate this concept, we removed the outlier (case 5_1) from the PCA and modeled reduced dimensional space based on 17 runs only. In fact, the outlier was suspected to have considerable baseline shift. The removed point then can be used as a test article to be compared against the reference standard.

Similar results were obtained as before, but the first two principal components explained 92% of the overall variation [92% (R^2), 88% (Q^2)]. The scores plot in Fig. 8 shows all 17 runs included in the analysis, and all of them are comfortably inside of 95% confidence ellipsoid and show random scatter.

To test articles against the reference standards, we have created a test set containing all 17 runs included in the PCA modeling and also the outlier, which was not included in the modeling. In other words, we used the entire 18 runs as test cases; we expect all 17 runs to be identified as belonging to the population of reference standards. On the other hand, case 5-1 (observation number 9) should be singled out as not belonging to the population.

Once the test set is defined, the new predicted scores for the test set can be obtained easily using the loadings of the training set (based on 17 runs only). Geometrically, this is equivalent to projecting the test data to the reduced dimension obtained from the training set (work set). Again using SIMCA, the distance from the PCA model

Table 6

Effects Table for Score2

	Coefficient Estimate	Standard Error	t for H_0 Coefficient = 0	Prob > |t|
Intercept	−1.25	29.26		
A DTT	5.22	29.26	0.18	0.8623
B Iodoacetate	42.69	29.26	1.46	0.1786
C Reduction *t*	−55.29	29.26	−1.89	0.0914
D-alkylation *t*	−56.00	29.26	−1.91	0.0879
E S/E	34.03	29.26	1.16	0.2747
F Digestion *T*	−16.48	29.26	−0.56	0.5870
G-Digestion *t*	−34.21	29.26	−1.17	0.2724

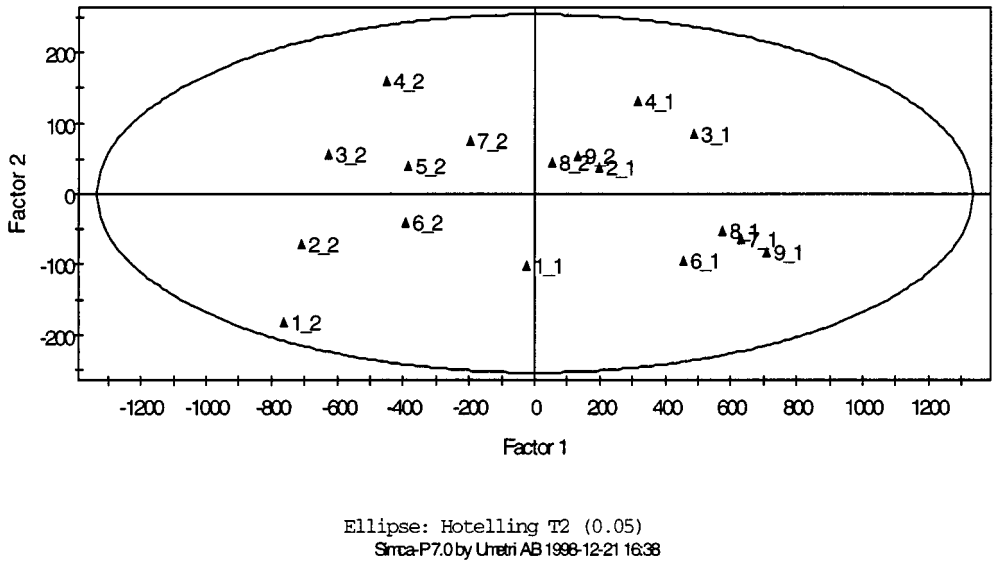

Figure 8. Scores plot without an outlier.

Table 7

Probability of an Article Belonging to Reference Standards

ObsNum	ObsName	Distance to the Model (*D*mod*X*)	Probability (*P*mod*X*)	Overall Distance (*T*square)
1	1_1	1.1029	0.28578	8.4204
2	1_2	0.6864	0.99041	11.4620
3	2_1	0.7738	0.94515	2.1523
4	2_2	0.9948	0.52198	6.8651
5	3_1	0.9475	0.63755	7.8370
6	3_2	0.6764	0.99252	7.4254
7	4_1	1.0782	0.33338	6.1792
8	4_2	0.6264	0.99817	8.6947
9	5_1	4.0413	0.00000	36.9230
10	5_2	1.1685	0.18179	1.8754
11	6_1	0.9318	0.67515	9.4289
12	6_2	0.9835	0.54953	5.6292
13	7_1	1.0580	0.37567	4.7605
14	7_2	1.3575	0.03768	2.3674
15	8_1	1.0616	0.36788	2.4390
16	8_2	0.7081	0.98421	5.0193
17	9_1	0.9994	0.51081	8.9576
18	9_2	1.4224	0.02055	2.4861

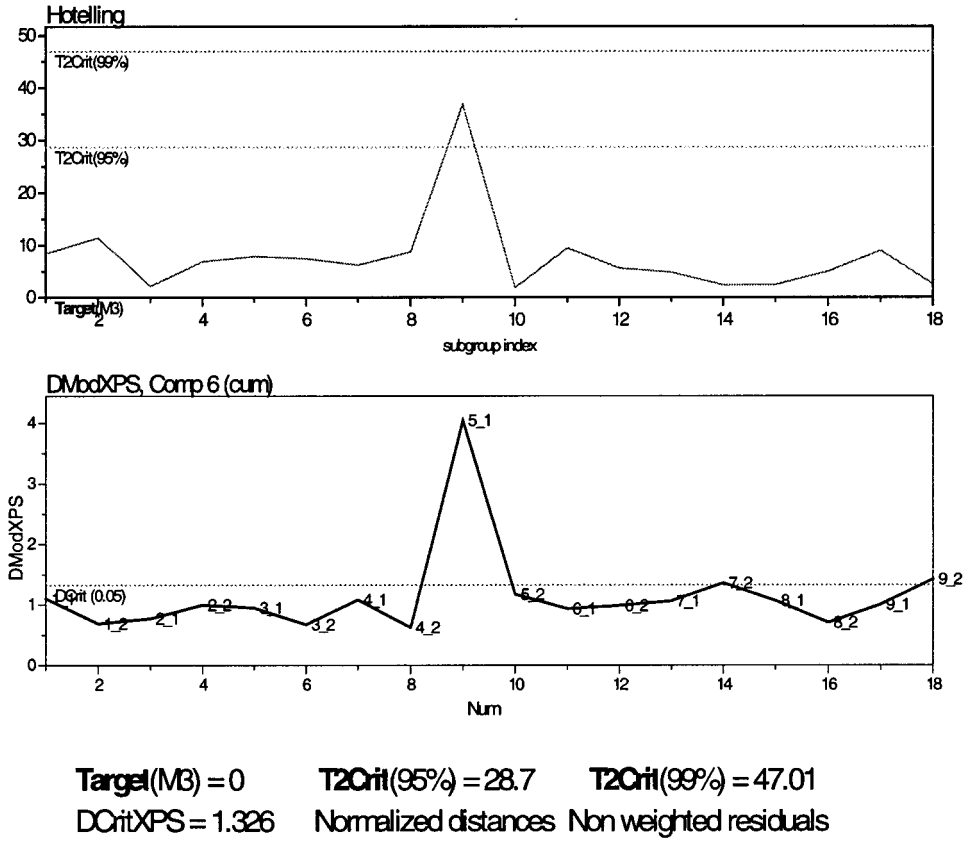

Figure 9. Control chart for identity of peptide maps.

(*D*mod*X*), the probability of the case belonging to the model (*P*mod*X*), and the overall distance based on all predicted scores in the reduced dimensions (*T*square) were calculated.

Note that, in Table 7, case 5_1 has an unusually large *D*mod*X* (4.04) and *T*square (36.9) and hence has a very small *P*mod*X* (0.000) value, which is an indication that this case is an outlier. These numbers can be depicted graphically, as in a statistical process control chart. Figure 9 shows an implementation in SIMCA that clearly shows that observation 9 (case 5_1) lies well outside the 95% control limits.

CONCLUSIONS

We have demonstrated that the chemometrics approach to the identity of a drug compound based on HPLC peptide mapping provides a statistically sound methodology. The chemometrics approach is essentially a multivariate approach that includes two statistical methods: experimental design and PCA. However, statistical analysis does not distinguish chromatographic variations due to instrument from true variations in sample composition, but we are really interested in detecting these variations in sample composition. Hence, to make the approach mentioned in this paper more accurate for detection of the variation in sample composition, it may be necessary to eliminate the chromatographic variations. One way to do this is by alignment of chromatographic profiles as a prerequisite for fingerprinting methods (see, e.g., Ref. 11). Validation studies of peptide maps for which pretreatment of data was done to eliminate chromatographic variations will be reported elsewhere.

REFERENCES

1. E. R. Hoff and R. C. Chloupek, Meth. Enzymol., 271, 51–68 (1996).
2. D. Allen, R. Baffi, J. Bausch, J. Bongers, M. A. Costello, J. Dougherty, Jr., M. Federici, R. Garnick, S. Peterson,

R. Riggins, K. Sewerin, and J. Tuls, Biologicals, 24, 255–275 (1996).
3. K. Kannan, M. G. Mulkerrin, M. Zhang, R. Gray, T. Steinharter, K. Sewerin, R. Baffi, R. Harris, and C. Karuntilake, J. Pharm. Biomed. Anal., 16, 631–640 (1997).
4. M. W. Dong, Adv. Chromatogr., 32, 21–51 (1992).
5. M. W. Dong and A. D. Tran, J. Chromatogr., 499, 125–139 (1990).
6. R. Newman, J. Alberts, James, D. Anderson, K. Carner, C. Heard, F. Norton, R. Raab, M. Reff, S. Shuey, and N. Hanna, Biotechnology, 10, 1455–1460 (1992).
7. S. Wold, K. Esbensen, and P. Geladi, Principal component analysis, Chemometrics Intelligent Lab. Sys., 2, 37–52 (1987).
8. SIMCA, SIMCA-P for Windows, version 3.0, Umetri AB, Umea, Sweden, 1996.
9. P. D. Haaland, in *Experimental Design in Biotechnology*, Marcel Dekker, New York, 1989, p. 37.
10. Design-Expert Version 5.0.7, 1997 Stat-Ease Corporation, Minneapolis, MN.
11. G. Malmquist and R. Danielsson, J. Chromatogr., 687, 71–88 (1994).