SURFACE AND INTERFACE ANALYSIS
Surf. Interface Anal. 2007; **39**: 427–433
Published online 23 February 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sia.2539

Surface and Interface Analysis

Adsorption of genetically engineered proteins studied by time-of-flight secondary ion mass spectrometry (TOF-SIMS). Part B: hierarchical cluster analysis (HCA)

Noriaki Suzuki, Mehmet Sarikaya and Fumio S. Ohuchi*

Department of Materials Science and Engineering, University of Washington, BOX 352120, Seattle, WA 98195-2120, USA

Received 16 July 2006; Revised 13 November 2006; Accepted 17 November 2006

In Part A, we adopted principal component analysis (PCA) for the analysis of TOF-SIMS data to assess the binding specificity of GBP-1 to metallic Au, Ag and Pd. Within a given set of data, PCA aids in the interpretation of the TOF-SIMS spectra by capitalizing on the differences from one spectrum to another. In Part B, we introduce another multivariate statistical method called 'hierarchical cluster analysis (HCA)', where visualization of the similarity and difference in data is readily observed, from which a variety of adsorption conditions of GBP-1 were characterized. Copyright © 2007 John Wiley & Sons, Ltd.

KEYWORDS: TOF-SIMS; HCA; PCA; genetically engineered proteins; GEPIs; GBP

INTRODUCTION

In Part A of this work, time-of-flight secondary ion mass spectroscopy (TOF-SIMS) was used to study the molecular interactions of one of the genetically engineered proteins for inorganics (GEPIs), specifically a gold binding protein (GBP-1), with metallic Au, Ag and Pd substrates. Principal component analysis (PCA) was successfully used to investigate characteristics of the GBP-1/substrate interactions using the fragmentation patterns. Data were analyzed to examine the cross specificity among the samples, and to identify the domains of the amino acid sequence that account for interactions with inorganic surfaces.

PCA reduces the multidimensional aspect of the TOF-SIMS spectra into two or three dimensions; therefore, the differences between the spectra and their cause can be readily recognized. PCA is mathematically matrix dependent, and it tries to identify the greatest variance within a given data set. If the data matrix includes groups of samples with some factors exceptionally different from others, score plots split into two groups to represent clear separation between them, but it does not show the difference within the groups. Within a subcategory of each group produced by the data matrix examined, this score plot does not readily provide further information. In theory, this problem can be solved by excluding the exceptional set of data and re-forming a new data set for another PCA process. To completely identify all the sample relations by PCA, a series of repetitive processes to reformulate a proper matrix are necessary. The unsupervised nature of the analysis results in significant overlap between spectra from different sets of the database, which can make data visualization, interpretation and classification difficult.[1–3]

In the case of GBP-1, we are fortunate to have various information available from various different sets of experiments. For example, we know that characteristic behavior toward certain inorganic materials influences the number of repeating units for GBP-1 and its conformation, etc. However, such information is not readily available for most of the proteins, which require extensive cross-examinations to learn the binding site of the protein. In practice, once the data matrix is produced, the next process to select the samples is totally up to the observer, and including one's intention to some extent is not totally avoidable. This 'human bias' may lead to misinterpretation of the data and subjective analysis of the results, particularly when no prior knowledge of the protein is available. Clearly, as the number of samples increases, the influence of bias becomes more significant, and verification of an objective view for PCA sample selection without human bias becomes more crucial, especially with limited prior information about the samples.

In Part B, we apply a multivariate statistical method called 'hierarchical cluster analysis (HCA)' to TOF-SIMS spectra to quantitatively determine the degree of 'similarity' and 'dissimilarity' among the samples. HCA is an unsupervised simple way of grouping the available data based on their similarities of selected variables.[4–11] In addition, HCA can be used to visualize the structure of the entire data set. This paper describes how HCA is applied to the data set from GBP-1 and discusses the structure of the data in terms of 'similarity and dissimilarity'.

*Correspondence to: Fumio S. Ohuchi, Department of Materials Science and Engineering, University of Washington, Seattle, WA 98195-2120, USA. E-mail: ohuchi@u.washington.edu

Surface and Interface Analysis

EXPERIMENTAL

Details about sample preparation, TOF-SIMS data acquisition and data normalization for the statistical analysis were described in Part A.[12] As a new set of the data in addition to the samples described in Part A, we include 'degraded' 3R : GBP-1 and 'denatured' GBP-1 into the data matrix in the present analysis. For denatured samples, the temperature is purposely set at 80 °C in an attempt to change the conformation of peptide upon adsorption. This is denoted as 'DN' throughout the paper. A list of samples investigated in this paper is summarized in Table 1. HCA technique is carried out without further standardizing the data, and the dendrograms are obtained using MATLAB.

Hierarchical cluster analysis (HCA)

'Cluster analysis' is based on the similarity of data in terms of a 'distance,' where a given specimen is taken as a point in the multidimensional space defined by the variables selected. The distance between this point and all other points is then calculated through a process called 'proximity matrix' to evaluate the similarity for all the specimens studied. The most frequently used method to calculate the distance is called 'Euclidean distance'. The Euclidean distance between objects X_i and X_j, is expressed as

$$D_{ij} = \left(\sum_{k=1}^{t} (X_{ik} - X_{jk})^2 \right)^{1/2}$$

where the number of variables is t. From the Euclidean distance, the matrix of proximity is obtained as follows:

$$S_{ij} = 1 - D_{ij}/\mathrm{MAX}\ (D_{ij})$$

where MAX (D_{ij}) is the longest interpoint distance. This function gives a combination of the most unlike set of the objects as $S = 0$, whereas the identical combination of the objects is $S = 1$. Once this proximity matrix is produced, an observer can scan for the smaller value in the proximity matrix. The corresponding cases are coupled together and considered as a single cluster. In stepwise manner, all the cases are clustered until one cluster includes all the cases.

Table 1. List of GBP-1 samples prepared

Sample	Substrate	Condition	No. of samples prepared	Total no. of scans
3R : GBP-1	Au	Room temp.	2	13
3R : GBP-1	Ag	Room temp.	3	9
3R : GBP-1	Pd	Room temp.	2	13
3R : GBP-1	Au	Degraded	1	4
3R : GBP-1	Au	Denatured	2	13
3R : GBP-1	Ag	Denatured	3	16
1R : GBP-1	Au	Room temp.	1	4
1R : GBP-1	Ag	Room temp.	2	10
1R : GBP-1	Pd	Room temp.	1	4
1R : GBP-1	Ag	Denatured	1	3
			Total	89

This leads to the construction of a two-dimensional diagram, known as a 'dendrogram'. Therefore, HCA transforms the complex data set and enables a visual overview of its characteristics in two dimensions.

Several mathematical methods allow grouping in a multidimensional space. Grouping produced by an agglomerative method is irrevocable by an algorithm. Thus, once introduced, defects in clusters cannot be repaired and care must be taken to form the data matrix. Some of the widely used agglomerative clustering methods are single linkage (the nearest-neighbor method), complete linkage (the furthest-neighbor method), average linkage, centroid linkage and Ward's error sum-of-square method. (Details of methods above are discussed elsewhere).[13,14]

Among all linkage methods available, one must choose the appropriate solution for the best representation of the original data in a dendrogram. One of the linkage methods, Ward's linkage, is a technique to minimize the loss of data as it forms consecutive links, and the Ward's linkage tends to form smaller clusters in the dendrogram.[15] The mathematical nature of Ward's linkage increases the ease of interpretation of that dendrogram because of distinctive subclusters. A large number of the samples must be examined for GBP-1; therefore, Ward's linkage is preferred. The following discussion is therefore based on the dendrogram with Ward's linkage.

RESULTS

Major clusters in the dendrogram

The dendrogram constructed by the Ward's linkage method is shown in Fig. 1. It reveals that the dendrogram consists of three majors clusters (Cluster (I), (II) and (III)), suggesting three distinctive fragmentation patterns within the samples examined. The degree of similarity/dissimilarity for those clusters is obtained from the dendrogram. On the basis of the Euclidean distance on the Y-axis, we can conclude that the characteristics of samples in Cluster (III) are significantly different from those in Cluster (I) and Cluster (II).

It was found that most of the samples classified into Clusters (I) and (II) are all denatured proteins, and the Clusters (I) and (II) are found at the first and second greatest separation in the dendrogram in Fig. 1. This suggests the dendrogram is dominated by the effect of denaturing; in other words, GBP-1 proteins are susceptible to heat, and therefore they are easily denatured when temperature is applied.

Consequently, the fragmentation patterns of SIMS are significantly altered because of the change of the conformation. Difference in interaction upon adsorption also influences the resultant fragmentation patterns. Although the cause of this difference cannot be determined by HCA, this dendrogram clearly indicated that GBP-1 is quite susceptible to denaturing by heat. To further look into the nature of GBP-1 adsorption, we studied the characteristics of each cluster.

Clusters (I) and (II)

Dendrograms for the Clusters (I) and (II) are separately shown in Figs 2 and 3, respectively. Most of the samples classified in Clusters (I) and (II) are from the denatured

DOI: 10.1002/sia

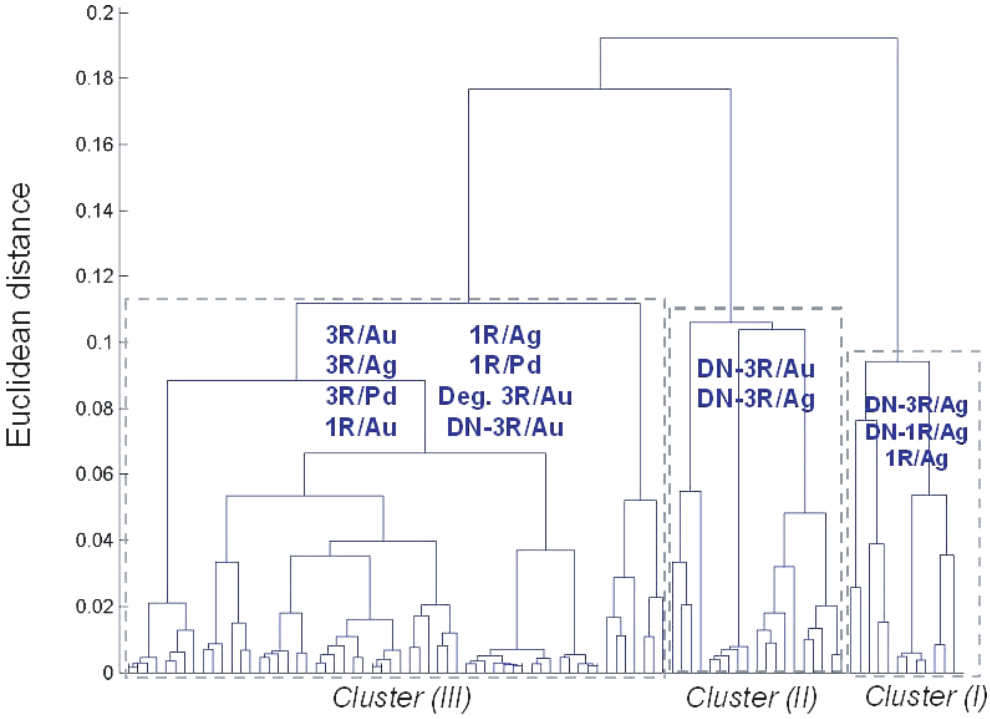

Figure 1. Dendrogram for all the samples studied by Ward's linkage. This figure is available in colour online at www.interscience.wiley.com/journal/sia.

GBP-1 regardless of the type of substrate used; the groups of samples for DN-1R/Ag and DN-3R/Ag are in Cluster (I), in which the Euclidean distance among those samples boxed by dotted lines takes similar values, indicating that the variations of fragmentation patterns from TOF-SIMS spectra for those samples are quite similar. In Cluster (II), all the samples are from triply repeated GBP-1 (3R : GBP-1), implying that the denatured 3R : GBP-1 at least behaves similarly regardless of the type of substrates used.

Careful examination of the type of samples belonging to the Clusters (I) and (II) revealed that DN-3R/Ag samples are occupied in both clusters. This implies that the denatured 3R : GBP-1 forms several distinctly different fragmentation patterns. In general, when proteins are denatured, their conformations change randomly, leading to a loss of the specific affinity toward the inorganic substrates. Contact to the substrate becomes nonspecific; therefore, the final conformation of the denatured proteins differs from sample to sample, even if the adsorption condition is identical. Appearance of the denatured GBP-1 in different clusters is consistent.

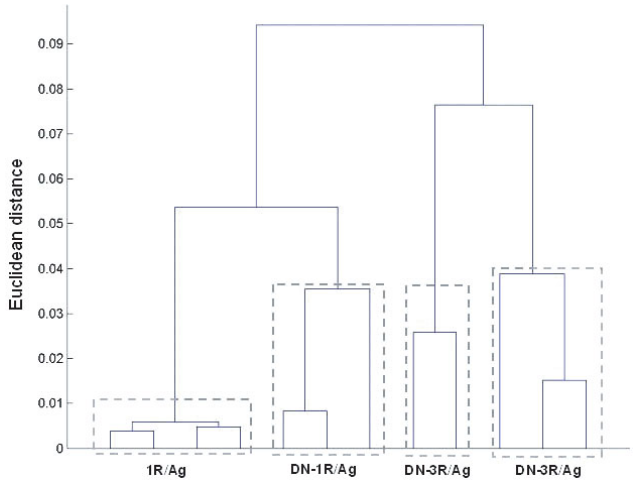

Figure 2. Dendrogram by Ward's linkage for Cluster (I). This figure is available in colour online at www.interscience.wiley.com/journal/sia.

The samples from 1R/Ag (adsorbed at room temperature) found in Cluster (I) are merged with the DN-1R/Ag group in the dendrogram. This indicates that the fragmentation patterns derived from 1R/Ag and DN-1R/Ag are similar. Since the 1R/Ag is the only nondenatured type among those found in Clusters (I) and (II), we concluded that this particular set of the samples 1R/Ag showed fragmentation patterns similar to the denatured samples, implying that adsorption of 1R on Ag must be either random or denatured for some reason during the adsorption process. As will be shown later, some of the 1R/Ag samples belong to Cluster (III).

Another set of the example that behaves differently in Cluster (III) is Deg-3R/Au (degraded 3R : GBP-1 adsorbed on Au), which does not fall into the same subgroup of the 'original' 3R : GBP-1 on Au. The value of proximity between 1R : GBP-1 on Au and degraded 3R : GBP-1 on Au ($S_{ij} = 0.85$) indicates that the fragmentation patterns of these two sets are highly similar. Since degraded 3R : GBP-1 had been highly ruptured during the degradation process, degradation shortened most of the proteins, resulting in the formation of the sequence similar to 1R : GBP-1.

Cluster (III)

Next, Fig. 4 provides a close-up view of Cluster (III). This consists of most of the samples adsorbed at room

DOI: 10.1002/sia

Surface and Interface Analysis

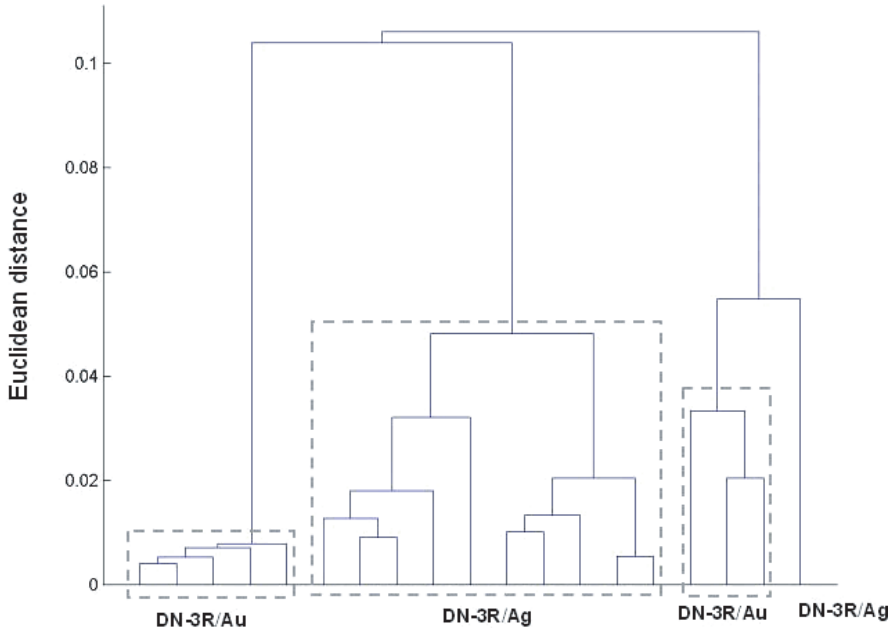

Figure 3. Dendrogram by Ward's linkage for Cluster (II). This figure is available in colour online at www.interscience.wiley.com/journal/sia.

temperature, and a few denatured sets. It is clear that each set of samples forms a subset of clusters, and their Euclidean distances are similar. The structure of the dendrogram reveals that the sample sets like 1R/Pd, 3R/Ag, 3R/Au and 1R/Ag are merged into one type of cluster (boxed in the red line) and further merged with a cluster consisting of 3R/Pd and 3R/Ag (boxed in the blue line). Another cluster contains two subsets of clusters consisting of 3R/Pd and DN-3R/Au (boxed in the green line). These two hierarchical clusters are merged into a larger one. A possible reason for a denatured sample (DN-3R/Au) to be included in the green box is the randomness of the denaturing characteristics discussed previously. The conformational change of the denatured protein is random, and therefore fragmentation patterns from DN-3R/Au samples most likely happen to be similar to those from 3R/Pd.

The advantage of using HCA is its ability to provide information about similarity and dissimilarity. Therefore, the question raised can be answered by simply looking at the location of the primary clusters in the dendrogram. In Fig. 4, one of the small clusters is DN-3R/Au, which appears in the green-lined box. This cluster merges into a bigger subcluster together with another cluster, 3R/Pd, at around 0.04 in Euclidean distance. This suggests that the TOF-SIMS fragmentation patterns for DN-3R/Au and 3R/Pd are at least similar. In other words, conformation of the denatured 3R-GBP as a result of the interaction with Au surface is similar to that seen in 3R-GBP adsorbed on Pd surface; however, the degree of conformational change by denaturing is not as significant as in samples found in Clusters (I) and (II) as discussed in the previous section.

Primary clusters from 3R/Au, 3R/Ag, 3R/Pd, 1R/Ag and 1R/Pd appeared several times in different sets of the subcluster groups shown in Fig. 4. For example, one of the 3R/Au clusters is seen in the red-boxed cluster and the same cluster is also found on far right-hand side of the dendrogram. Similarly, 3R/Pd is present in the blue-boxed cluster, and another 3R/Pd appears as a primary cluster form; however, they are quite different in character in terms of Euclidean distances. This means that the first greatest variance cannot easily separate these samples on the basis of the type of substrate, which is consistent with the PCA analysis described in Part A.

Another set of the example that behaves differently in Cluster (III) is Deg-3R/Au (degraded 3R : GBP-1 adsorbed on Au), which does not fall into the same subgroup of the 'original' 3R : GBP-1 on Au. The value of proximity between 1R : GBP-1 on Au and degraded 3R : GBP-1 on Au ($S_{ij} = 0.85$) indicates that the fragmentation patterns of these two sets appear to be highly similar. Since degraded 3R : GBP-1 likely had been highly ruptured during the degradation process, degradation most likely shortened this protein, resulting in the formation of the sequence similar to 1R : GBP-1.

Comparison of HCA and PCA

On the basis of the dendrogram, HCA clearly suggests that GBP-1 proteins are extremely susceptible to heat and easily denatured when temperature is applied, causing the alteration of fragmentation patterns accordingly. To justify the coherency of these findings, PCA is carried out for the comparison as well as the cause determination for its spectrum change in denaturing. For PCA data matrix construction, data are included from all samples examined previously for HCA. It includes various conditions, such as Au, Ag and Pd substrates, and also 1R and 3R : GBP, adsorption at room temperature and high temperature for denaturation. Figure 5 shows the score plot of PC1 and PC2.

The score plot on PC1 shown in Fig. 5(a) suggests that the most significant difference is found between denatured 3R : GBP on Ag and denatured 1R : GBP on Ag. It should be noted that these two samples were previously found in Cluster (I) by HCA. One may notice that denatured 3R : GBP

DOI: 10.1002/sia

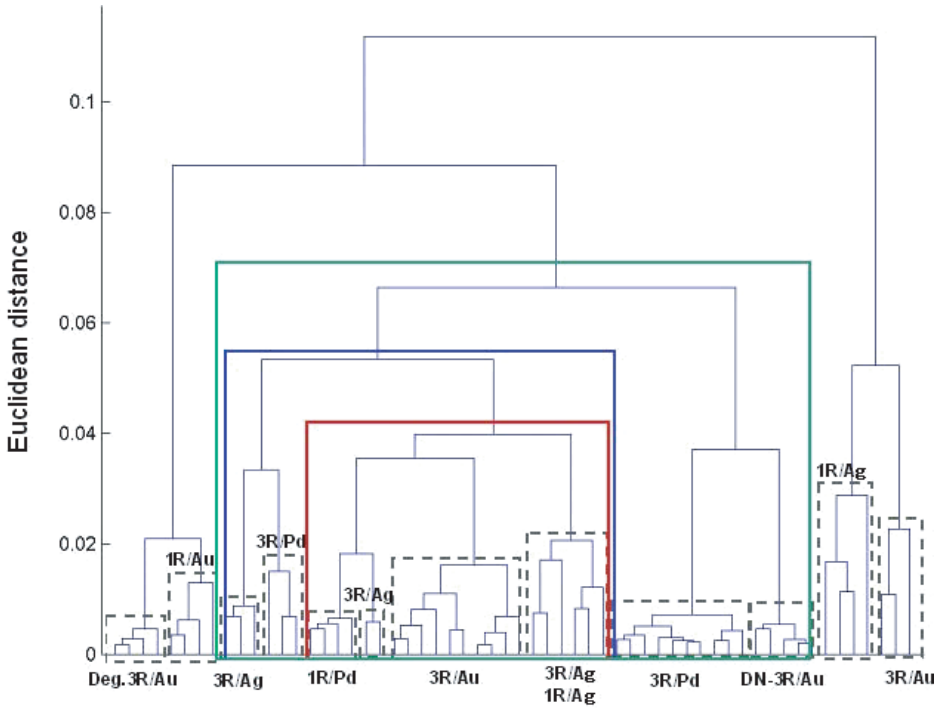

Figure 4. Dendrogram by Ward's linkage for Cluster (III). This figure is available in colour online at www.interscience.wiley.com/journal/sia.

on Ag samples have scores on both the positive and the negative side. This result is expected because the protein is purposely denatured and denaturing occurs randomly, resulting in the large deviation in fragmentation pattern in the spectra, which suggests instability. As PC1 explains only 33% of the original data, we also investigated scores on PC2 shown in Fig. 5(b).

PC2 in the score plot clearly separates denatured samples from samples adsorbed at room temperature. The positions of denatured samples were significantly scattered in the score plot. Since these are adsorbed onto the substrate randomly, this deviation is again expected. From the scores on PC1 and PC2, we can conclude that GBP-1 is highly susceptible to denaturing by applying the temperature upon adsorption. This finding is consistent with that of HCA, as reported previously. Once GBP-1 is denatured, regardless of number of repeating unit, the fragmentation patterns are totally different from the samples adsorbed at room temperature. This finding reinforces the issues discussed in Part A that the fragmentation pattern indeed originates from the interaction of the protein. The effect of denaturation is revealed by the corresponding loading plot shown in Fig. 6, where negative loading corresponds to the denatured proteins.

Loadings on PC2 show many fragments of very high value either in the positive or negative side. The dominant peak in the negative side corresponding to denatured samples is $m/z = 30$. This fragment is integral to the six amino acids G, H, I, K, M and S. It is rather difficult to determine the individual contributions. However, this specific fragment, and other shared fragments (such as $m/z = 56$), in the larger negative values suggests that many amino acids in the sequence are influenced by the denaturing effect.

Owing to the inability to determine the individual amino acid contribution from loadings in this case, it is possible that most of the amino acids are strongly influenced by denaturing. Therefore, the stable motif found previously in Part A may no longer exist. This simply implies that the denaturing strongly influences the loss of the functionality to adsorb on substrates in an ordered manner. This result is consistent with the previous discussion.

DISCUSSION

This paper described the reanalysis of TOF-SIMS data by HCA to visualize, interpret and classify the samples. The dendrogram is separated into three major groups on the basis of the adsorption temperature. It indicates the resistance of GBP-1 to denaturing by heat, and also shows its randomness in conformation. This suggests that nondenatured samples have distinctive fragmentation patterns compared to denatured samples. To further investigate the differences within the samples in the same group, we examined the smaller clusters in more detail.

As shown in Fig. 4, Cluster (III) comprised two small clusters of 3R : GBP-1/Au, three small clusters of 3R : GBP-1/Ag and two small clusters of 3R : GBP-1/Pd, which are all scattered within this dendrogram, and no definite correlation between positions of each cluster and type of sample is apparent. All the samples found in this Cluster show similar fragmentation patterns in their TOF-SIMS spectra, and the slight fragmentation difference in each sample most likely causes the scattered position of each subcluster. However, this information in terms of fragmentation similarities was not easily obtained from PCA as its mathematical nature meant that only the largest

DOI: 10.1002/sia

Surface and Interface Analysis

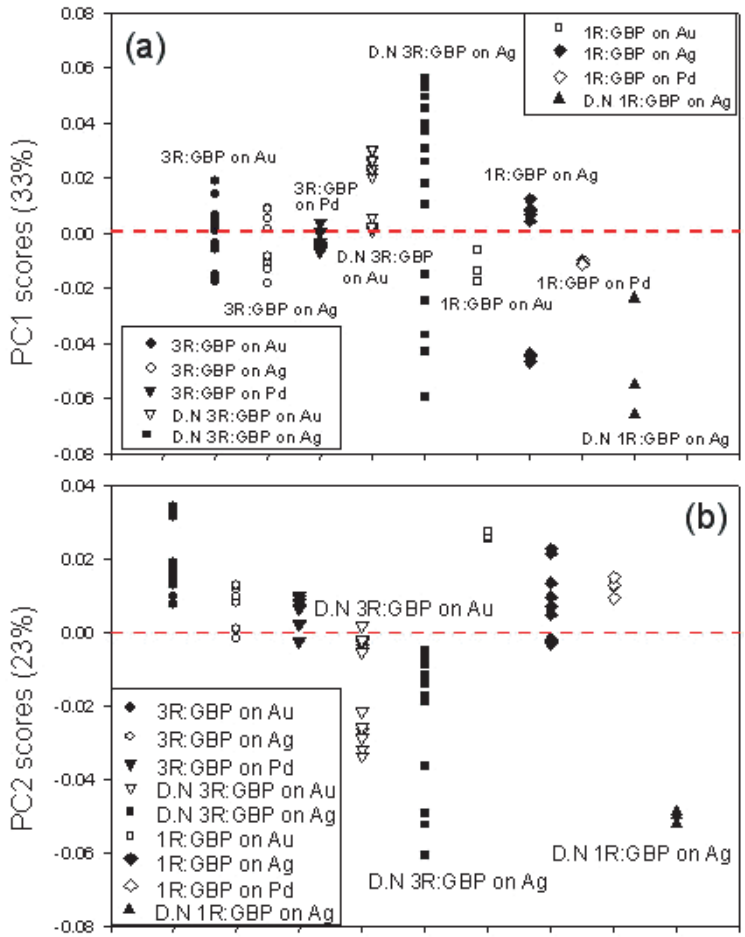

Figure 5. Scores for all the samples examined on the basis of (a) PC1 and (b) PC2. This figure is available in colour online at www.interscience.wiley.com/journal/sia.

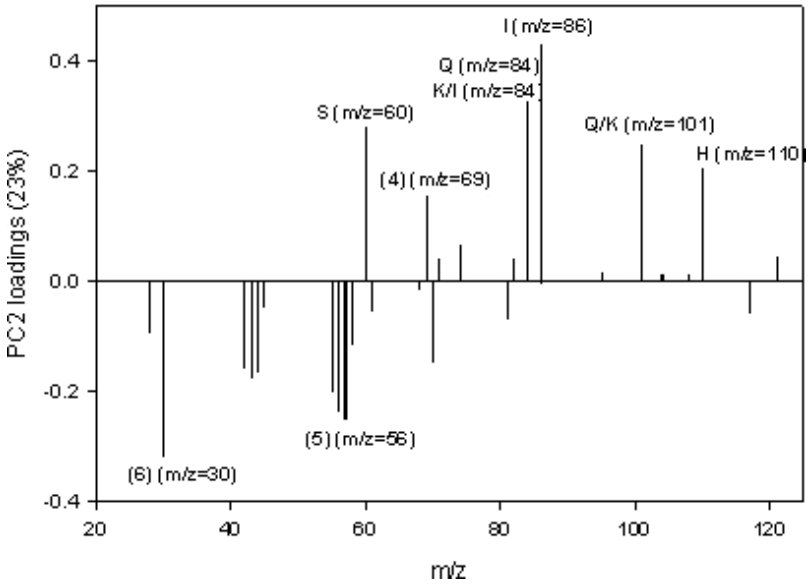

Figure 6. Loadings on the PC2 for all the samples examined.

difference within a given data set was extracted. Thus, we suggest the possible use of HCA as a complementary analysis process with PCA.

A good example is the comparison of 3R:GBP-1/Au *versus* 3R:GBP-1/Ag, which was also discussed in Part A. According to Fig. 4, one of two 3R:GBP-1/Au subclusters found on the right side of the dendrogram differs from the group consisting of three 3R:GBP-1/Ag subclusters and the other 3R:GBP-1/Au subcluster found in the blue dotted box. If an observer constructs the data matrix and intends to

DOI: 10.1002/sia

investigate the interaction of 3R : GBP-1 on the basis of the type of substrate by PCA, the first greatest variance will not separate the samples on the basis of the type of substrate. To satisfy our purpose, we recommend using several PCs for analysis. Indeed, this is consistent with the results reported in Part A.

These findings suggest that an unsupervised HCA can: (i) provide an overview of the fundamental characteristics of each sample at a glance and determine the grouping on the basis of the most influential conditions; (ii) be coupled in a complementary way with PCA to supplement the information in similarity; (iii) exclude inappropriate samples, such as artifacts, from the data set for further PCA and (iv) be used as a good indication for the extent of considering principal components in PCA.

CONCLUSIONS

In part A, we concluded that TOF-SIMS can provide the information on (i) the type of amino acids likely involved in the interactions and their motif that account for specific affinity toward noble metals and (ii) the specificity of protein toward inorganic materials. These assessments were successful owing to the extensive previous studies and the prior knowledge of the intrinsic behavior of GBP-1. In addition, as the number of samples to be analyzed increases, creation of the data matrix for PCA requires extra care for proper sample selection. Here in Part B, we introduced another multivariate analysis, HCA. This technique allows us to visualize the relation of a number of samples on the basis of similarity and dissimilarity of the fragmentation pattern. HCA at a glance revealed the heat-susceptive nature of GBP-1. Therefore, in addition to (i) and (ii), HCA is able to provide (iii) conformational recognition based on the fragmentation pattern. In previous discussion, denatured 3R : GBP-1 suggested a significant alteration of protein conformation, and therefore the suggested binding motif is no longer responsible for the binding.

The relation of subclusters found in the dendrogram is consistent with what we observed in PCA, suggesting the strong possibility that we can use HCA as a preliminary data screening prior to PCA. To further refine this protocol, however, there are pending issues: (i) correlation of amino acid fragmentation yield with binding strength; (ii) correlation of amino acid fragmentation yield with type of interaction; (iii) determination of individual amino acid and its contribution from the shared fragment and (iv) deviation of certain fragment yield to another fragment from the same amino acid.

Acknowledgements

This project was supported by the UW-DURINT program (Defense University Research Initiative on Nanotechnology, PM: Dr. Robert Campbell) through the US-Army Research Office (Grant No. DAAD19-01-1-04999) and partially funded from NSF MRSE program (Grant # DMR520567). The authors also acknowledge support from the National ESCA and Surface Analysis Center for Biomedical Problems (NESAC/BiO), funded by NIBIB, Grant # EB002027. The authors also acknowledge Professors David Castner and Lara Gamble for their assistance in interpretations of TOF-SIMS data.

REFERENCES

1. Yu P. *J. Agric. Food Chem.* 2005; **53**: 7115.
2. Oust A, Moretro T, Kirschner C, Narvhus JA, Kohler A. *J. Microbiol. Methods* 2004; **59**: 149.
3. Beckonert O, Bollard ME, Ebbels TMD. *Anal. Chim. Acta* 2003; **490**: 3.
4. Honorio E, Lirio A, Mateus C, Barros PP, Valente M. *CASEMIX* 2000; **2**: 39.
5. Eisen MB, Spellman PT, Brown PO, Botstein D. *Proc. Natl. Acad. Sci. U.S.A.* 1998; **95**: 14863.
6. Ferreira MMC, Faria CG, Paes ET. *Chemom. Intell. Lab. Syst.* 1999; **47**: 289.
7. Mantas A, Deretey E, Ferretti FH, Estrada MR, Csizmadia IG. *J. Mol. Struct. THEOCHEM* 2000; **504**: 171.
8. Kowalski BR, Bender CF. *J. Am. Chem. Soc.* 1972; **94**(16): 5632.
9. Shannon W, Culverhouse R, Duncan J. *Pharmacogenomics* 2003; **4**(1): 41.
10. Pripp AH, Rehman SU, Mcsweeney PLH, Fox PF. *Int. Dairy J.* 1999; **9**: 473.
11. Felicissimo MP, Peixoto JLS, Thoman R, Azioune A, Pireaux JJ, Houssiau L, Filho UPR. *Philos. Mag.* 2004; **84**(32): 3483.
12. Suzuki N, Gamble L, Tamerler C, Sarikaya M, Castner DG, Ohuchi FS. *Surf. Interface Anal.* 2007; **39**: 419.
13. Dillon WR, Goldstein M. *Multivariate Analysis Methods and Applications*. Wiley: New York, 1984.
14. Anderberg MR. *Cluster Analysis for Applications*. Academic Press: New York, London, 1973.
15. Ward JH. *J. Am. Stat. Assoc.* 1963; **58**: 236.

DOI: 10.1002/sia