LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-202928

Imaging of Isotopically Enhanced Molecular Targeting Agents Final Report

J.N. Quong

March 16, 2004

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

UCRL- TR-202928

Imaging of Isotopically Enhanced Molecular Targeting Agents Final Report

J.N. Quong

16 March 2004

Imaging of Isotopically Enhanced Molecular Targeting Agents Final Report
Judy N. Quong

Summary

The goal of this project is to develop experimental and computational protocols to use SIMS to image the chemical composition of biological samples, focusing on optimizing sample preparation protocols and developing multivariate data analysis methods. Our results on sample preparation, molecular imaging, and multivariate analysis have been presented at several meeting abstracts (UCRL151797ABS, UCRL151797ABSREV1, UCRL151426ABS, UCRL201277, UCRL154757). A refereed paper describing our results for sample preparation and molecular imaging of various endogenous biomolecules as well as the mutagen PhIP has been accepted for publication (UCRL-JC-151797). We are also preparing two additional papers describing our multivariate analysis methods to analyze spectral data. As these papers have not been submitted, their content is included in this final report.

Introduction/Motivation

Time Of-Flight Secondary Ion Mass Spectrometry (TOF- or static-SIMS) is a surface analysis technique that provides chemical specificity and imaging capability. SIMS has been used to characterize biomaterial surfaces, molecules in tissues, and proteins (Kempson 2003, Roddy 2003, Wagner 2003). The time-of-flight mass analyzer generates complex spectra containing large numbers of peaks, primarily those originating from proteins that constitute most of a dry cell's mass. Proteins are long unbranched chains composed of only 20 different amino acids. The complexity lies in the large number of combinations of these 20 amino acids present in cells. Adding to the complexity is the fact that in general, a given biomolecules can be ionized not only as single 'parent' ions, but also as multiple 'daughter' ionized fragments. Thus, in general, there is an absence of unique peaks for different biological samples.

Identification of signatures from mass spectra in conjunction with genomic data is being developed for applications such as disease diagnostics and characterization as well as proteomic profiling (Diamandis 2004, Nishizuka 2003). One of the difficulties with mass spectra is the complex multidimensionality of the dataset. These data are over determined; that is, the number of variables is greater than the number of samples. Algorithms have been used to analyze these data to demonstrate that a pattern consisting of several peaks (from tens to thousands) is sufficient to differentiate between two groups using genetic algorithms combined with cluster analysis (Petricoin 2003). Principal component analysis (PCA) has also been used to interpret TOF-SIMS spectra recorded from proteins adsorbed onto model surfaces (Lhoest 2001). In these methods, a subset of m/z peaks are selected to represent a given protein sample based on the observed spectra and amino acid composition. Two supervised multivariate analysis techniques, discriminant principal component analysis and linear discriminant analysis, have been compared in their ability to distinguish spectra from adsorbed protein films (Wagner

2002). PCA has also been used to analyze TOF-SIMS images of organic monolayers of single and mixed phospholipids (Biesinger 2002).

Recently, algorithms to characterize underlying structure in complex gene expression data have been developed using 'metagene' patterns from DNA microarray experiments. In this method, singular value decomposition (SVD) is used to derive metagenes that are linear combinations of individual gene expression values that together constitute the metagene (West 2000). The metagenes have been shown to identify and classify cellular phenotypes based upon their gene expression values.

For this project, we used Singular Value Decomposition to calculate linear combinations of the original mass/charge peaks contained in an alternate coordinate system. We then selectively remove, or project out, independent m/z information to remove the 'commonality' contained in the spectral information to expose underlying differences in different sample spectra. We show that we are able to cluster spectra from homogenates of different cellular regions and cell types. A second application is to use SVD to derive 'metamasses' from TOF-SIMS spectra. We then identify those metamasses that contribute most to the intra-group separation using canonical analysis (CA). In order to relate these metamasses back to the original measured mass peaks, we again use canonical analysis to identify those mass peaks that contribute most to the intra-group separation. We find that SVD in combination with canonical analysis (CA) enabled the identification of a subset of m/z peaks that are responsible for giving maximal separation of different sample types. Using this method, for the samples we used, we find that a signature comprised of 35 m/z peaks is sufficient to categorize spectra from a variety of biological samples.

Data Analysis

TOF-SIMS spectra contain thousands of peaks. The challenge is identifying the peaks in the spectra that can be used as the markers of a complex signature. In order to reduce the dimensionality of the space spanned by the mass spectra, we first apply singular value decomposition (SVD) to the training dataset. The equation for the singular value decomposition of M is: $M = USV^T$ where U is an m x n matrix, S is an n x n diagonal matrix, and V^T is an n x n matrix. The columns of U are the left singular vectors and form an orthonormal basis for the mass spectra. Each single vector of U represents a linear combination of m/z peaks and this linear combination is referred to as a metamass. The resulting matrix V^T contains the right-singular vectors and represents the measurements of the metamass for each observation. By representing the data in this new coordinate system defined by the left singular vectors of the SVD, the dimensionality is reduced from n the number of m/z peaks to at most m, the number of samples. It is important to note, that this representation of the data is exact and no information is lost. The set of left singular vectors spans the identical space as the original data. Each of these new variables, which we call metamasses, is a linear combination of the original peak intensities. For the clustering application, we define M(m), the new data matrix defined by the removal of the m largest eigenvectors.

In order to identify signatures from the spectra, we then use a variant of canonical analysis on the reduced space to determine the separation between the different spectra. Canonical analysis only works in the case when $m < n - g$, where g is the number of distinct groups. In our case, after SVD, m=n, and g=4 so we cannot directly apply Canonical analysis. To get around this problem, we apply a selection procedure know as forward selection. Forward selection is an iterative procedure and finds the best set of 1, 2, 3, …, n-g metagenes that maximizes separation between the groups. The best set of metagenes for small sets, typically up to about 5, we perform the optimization by considering all possible combinations of metagenes and selecting the set that yields the best Wilke's parameter. For larger sets, the exhaustive search is too time consuming and we add a single metagene by considering the given group and the addition of a single metagene. We repeat this process until we have found n-g sets. One problem with this approach and with almost all multivariate approaches is that the problem of "cherry picking." Because we have as many degrees of freedom (variables) to choose from as observations, we can always find a set of n-g variables that will provide separation between groups. The challenge here is to find a set of variables that not only provides separation, but also is predictive. To help identify sets of genes that provide good separation, but are not statistically significant, we calculate the significance of each added variable during each step of the forward selection procedure. We then only choose those sets of metagenes that are significant at the level of 0.05.

Because the transformation from the mass spectra to metamasses is dependent upon the samples, we need to identify the particular peaks in the spectra that can be used as the components of a signature. To this end, we consider each metamass that has been identified in the first step and take only the mass peaks that contribute significantly to the metagene. For this, we take peaks where the coefficients from the left singular vector that makes up the metamass $c(ij)^2 > tol$, where i identifies the m/z peak in metamass j. This is repeated for each metamass yielding a list of potential m/z peaks for use in identifying the signature.

The final step is to apply the forward selection to the list of m/z peaks identified in the second step. The criteria we use to avoid any false positives are the same as in the first forward selection.

Clustering of Cell Homogenate Spectra Using Singular Value Decomposition

As described earlier, ions analyzed by TOF-SIMS are generated directly from the sample surface. The majority of cell mass is composed of proteins (polymers of only 20 different amino acids) comprise the majority of dry cell mass. Thus, TOF-SIMS mass spectra obtained from biological material, including cells, are very similar. Figure 1 shows typical mass spectra obtained from cell homogenates of the nuclear and particulate fractions of the MCF-7 breast cancer cell line. The similarity can be quantified by calculating the correlation coefficient between mass spectra obtained from homogenates from different cellular regions. Table 1 shows the correlation coefficients calculated between 5 spectra collected from cytosolic, nuclear, and particulate fractions. By definition, the correlation coefficient with any mass spectra with itself is 1.0 as seen

along the diagonal. The cytosolic and nuclear fractions are highly correlated within each group and between groups (correlation coefficient > 0.99) while the correlation between the particulate fraction and the other two fractions are much lower (correlation coefficient < 0.52).

The S matrix containing 30 eigenvalues was calculated from the data matrix containing 30 TOF-SIMS spectra (m/z 1-400) from cytosolic, nuclear, and particulate fractions. The presence of 30 non-zero eigenvalues indicates that the mass spectra are linearly independent. The SVD orders the eigenvalues high-to-low beginning at the upper left corner. For this dataset, the largest eigenvalue contains ~ 70% of the variance of the data (not shown). After removal of the contribution of the largest eigenvector by setting the corresponding eigenvalue to 0.0, we define the new data matrix M(m) and then calculate the correlation coefficients between the newly represented TOF-SIMS spectra contained in M(m). The results are shown in Table 2. The within-group correlation decreased to < 0.1 in some cases. However, *all* between-correlations were < 0.0, enabling classification of the spectra into a cytosolic, nuclear, and particulate fractions.

Identification of Signatures to Distinguish SIMS Spectra from Biological Samples

We analyzed 30 mass spectra from lysozyme, cytochrome c, and myoglobin samples identified two metamasses that gave the best separation along canonical axis 1 and 2 (not shown). However, we were unable to validate these m/z peaks due to day-to-day variability in spectra. We therefore collected an additional 188 spectra on four different days on different silicon wafer chips. We used 19 myoglobin spectra and 20 spectra each from cytochrome C, lysozyme, and insulin as the training set and the remaining spectra as the validation set. In this case rather than 2 m/z peaks, 35 peaks were required to separate the four protein spectra. Figure 2 shows the plot of the first two canonical axes of the training (solid circles) and validation (open circles) spectra. Most of the separation is along the first canonical axis. For the training data, all four proteins can be distinguished, primarily along the first canonical axis although the lysozyme and insulin spectra are very close together. When the validation spectra are overlaid onto the training set, the cytochrome C and myoglobin test set overlay the training data and are clearly separated; however, the lysozyme and insulin cannot be distinguished. For reasons we do not yet understand, there is a large amount of scatter in the insulin test data.

Summary

We developed protocols to measure ions generated from biological samples and measured the distribution of various biomolecules. PhIP distribution was localized to the plasma membrane, but development of data analysis methods is required for image interpretation. We have developed a method using singular value decomposition and a variant of canonical analysis to distinguish mass spectra from different samples including proteins, cell homogenates, and images.

References

Biesinger, M.C.; Paepegaey, P.Y.; McIntyre, N.S.; Harbottle, R.R.; Petersen, N.O. Principal component analysis of TOF-SIMS images of organic monolayers. Anal Chem. 2002 Nov 15;74(22):5711-6.

Diamandis, E.P. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: Opportunities and potential limitations. Mol Cell Proteomics. 2004 Jan 30.

Kempson IM, Skinnera WM, Kirkbride, PK. Calcium distributions in human hair by ToF-SIMS. Biochimica et Biophysica Acta 1624 (2003) 1– 5.

Lhoest, J.-B.; Wagner, M.S.; Tidwell, C.D.; Castner, D.G. J. Biomed. Mater. Res. 2001, 57, 432-440.

Nishzuka, S.; Chen, S.T.; Gwadry, F.G.; Alexander, J.; Major, S.M.; Scherf, U.; Reinhold, W.C.; Waltham, M.; Charbonaeau, L.; Young, L.; Bussey, K.J.; Kim, S.; Lababidi, S.; Lee, J.K.; Pittaluga, S.; Scudiero, D.A.; Sausville, E.A.; Munson, P.J.; Petricoin, E.F. 3rd; Liotta, L.A.; Hewitt, S.M.; Raffeld, M.; Weinstein, J.N. Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic, and tissue array profiling.Cancer Res. 2003 Sep 1;63(17):5243-50

Petricoin, E.F>; Ardekani, A.M.; Hitt, B.A.; Levine, P.J.; Fusaro, V.A.; Steinberg, S.M.; Mills, G.B.; Simone, C.; Fishman, D.A>; Kohn, E.C.; Liotta, L.A. Use of proteomic patterns in serum to identify ovarian cancer.
Lancet. 2002 Feb 16;359(9306):572-7

Roddy, T.P.; Cannon D.M. Jr; Ostrowski, S.G.; Ewing, A.G.; Winogra, N. Proton transfer in time-of-flight secondary ion mass spectrometry studies of frozen hydrated dipalmitoylphosphatidylcholine. Anal. Chem. 2003 Aug 15:75(16):4087-94.

Wagner, M.S.; Tyler, B.J.; Castner, D.G. Interpretation of Static Time-Of-Flight Secondary Ion Mass Spectra of Adsorbed Protein Films by Multivariate Pattern Recognition. Anal. Chem. 74(8):1824-35, 2002.

Wagner, M.S.; Shen, M.; Horbett, T.A.; Castner, D.G. Quantitative analysis of binary adsorbed protein films by time of flight secondary ion mass spectrometry. J Biomed Mater Res 2003 Jan 1:64A(1):1-11.

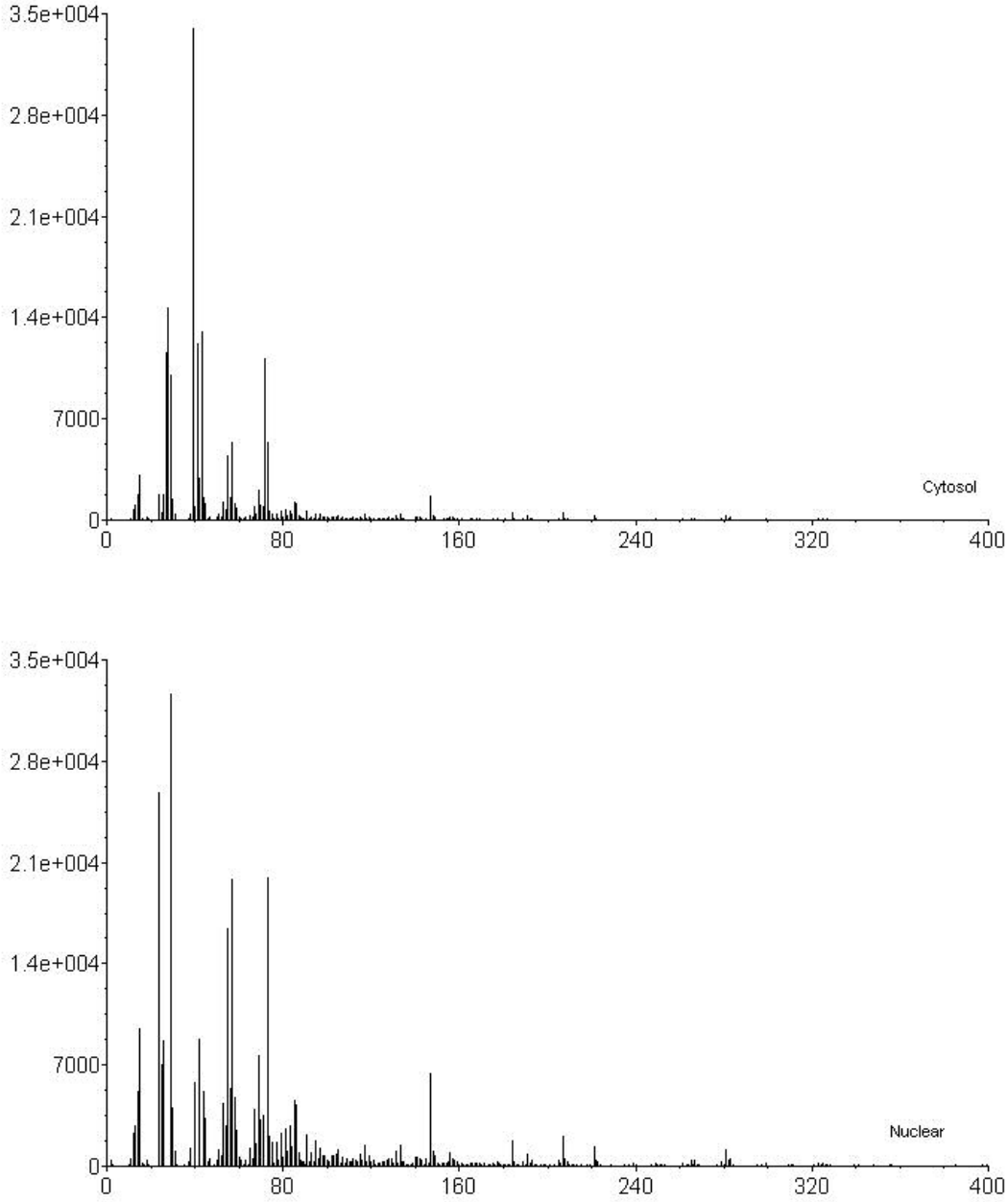

Figure 1. Typical mass spectra obtained from cell homogenates of the nuclear and particulate fractions of the MCF-7 breast cancer cell line.

	C1	C2	C3	C4	C5	N1	N2	N3	N4	N5	P1	P2	P3	P4	P5
C1	1														
C2	0.999	1													
C3	0.999	0.999	1												
C4	0.992	0.996	0.997	1											
C5	0.995	0.998	0.999	0.999	1										
N1	0.996	0.994	0.995	0.989	0.991	1									
N2	0.992	0.990	0.990	0.982	0.984	0.999	1								
N3	0.998	0.983	0.984	0.972	0.976	0.993	0.997	1							
N4	0.994	0.993	0.993	0.986	0.989	0.999	0.998	0.994	1						
N5	0.995	0.992	0.993	0.985	0.988	0.999	0.999	0.996	0.999	1					
P1	0.439	0.410	0.413	0.362	0.377	0.467	0.512	0.552	0.469	0.488	1				
P2	0.449	0.419	0.422	0.368	0.383	0.475	0.519	0.559	0.477	0.496	0.998	1			
P3	0.479	0.453	0.458	0.411	0.423	0.509	0.552	0.591	0.509	0.529	0.996	0.989	1		
P4	0.451	0.423	0.427	0.378	0.391	0.482	0.527	0.565	0.483	0.502	0.999	0.995	0.998	1	
P5	0.508	0.478	0.480	0.426	0.443	0.530	0.572	0.611	0.533	0.550	0.992	0.997	0.984	0.988	1

Table 1. Correlation matrix of raw TOF-SIMS data. A correlation value of 1 represents two identical samples and as the value decreases the differences between samples increases. The correlation between nuclear and cytosolic fractions renders them virtually indistinguishable. Only the membrane bound particulate fraction can be identified.

	C1	C2	C3	C4	C5	N1	N2	N3	N4	N5	P1	P2	P3	P4	P5
C1	1														
C2	0.865	1													
C3	0.764	0.982	1												
C4	0.459	0.828	0.888	1											
C5	0.587	0.905	0.948	0.986	1										
N1	-0.17	-0.10	-0.96	0.070	0.000	1									
N2	-0.72	-0.78	-0.77	-0.58	-0.67	0.681	1								
N3	-0.67	-0.84	-0.83	-0.80	-0.83	0.0583	0.628	1							
N4	-0.14	-0.15	-0.19	-0.07	-0.12	0.904	0.631	0.27	1						
N5	-0.44	-0.47	-0.48	-0.32	-0.40	0.880	0.867	0.480	0.923	1					
P1	-0.70	-0.93	-0.93	-0.90	-0.93	-0.230	0.543	0.798	-0.14	0.174	1				
P2	-0.66	-0.91	-0.93	-0.92	-0.94	-0.230	0.534	0.792	-0.13	0.174	0.997	1			
P3	-0.75	-0.93	-0.92	-0.87	-0.91	-0.240	0.535	0.787	-0.17	0.153	0.996	0.986	1		
P4	-0.73	-0.94	-0.94	-0.89	-0.93	-0.210	0.563	0.797	-0.13	0.190	0.999	0.993	0.998	1	
P5	-0.62	-0.90	-0.92	-0.91	-0.94	-0.240	0.512	0.790	-0.14	0.159	0.999	0.999	0.978	0.986	1

Table 2. Correlation matrix of data after SVD. A correlation value of 1 represents two identical samples and as the value decreases the differences between samples increases. The within-group correlation decreased to < 0.1 in some cases. However, *all* between-correlations were < 0.0, enabling classification of the spectra into a cytosolic, nuclear, and particulate fractions.

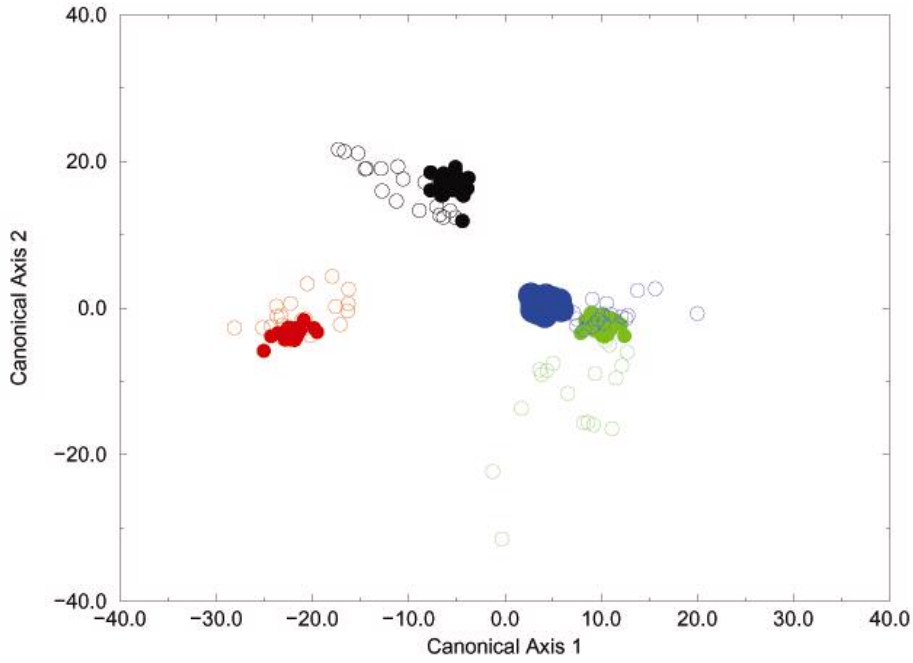

Figure 2. Myoglobin (blue), cytochrome C (red), lysozyme (blue), and insulin (green) spectra plotted along canonical axes 1 and 2. Solid circles are spectra from the training set. Open circles are spectra from the validation set.