

Grouping of Partially Methylated Alditol Acetates from their GC-EIMS spectra using Principal Component Analysis, non-parametric density estimation and k-Nearest Neighbor Classification

Arora, I.

Graduate Student,
Department of Computer Science,
San Diego State University

Valafar, F.

Associate Professor,
Department of Computer Science,
San Diego State University

Abstract – *Recent developments in identification of complex carbohydrates have indicated that pattern recognition techniques can effectively be applied to structural elucidation and categorization of these macromolecules. In this paper, we present one such classification methodology. Our technique utilizes data compression and feature selection using discrete Karhunen-Loeve expansion, also known as Principal Component Analysis (PCA). Multi-dimensional density estimation and k-Nearest neighbor classifier was then applied to the feature values computed by PCA. A cluster based approach using Bayes' statistics was subsequently adopted to divide the problem space into cascading linearly separable sub-spaces. The training set consisted of the gas chromatography-electron impact mass spectroscopy (GC-EIMS) spectra of 363 partially Methylated Alditol Acetate (PMAA) of monosaccharide residues. Our system was able to separate 8 of the 11 groups of PMAAs with high accuracy. A comparative discussion of our results and those of similar studies has also been offered and conclusions have been drawn.*

Keywords: Partially methylated alditol acetates, Principal Component Analysis, k-NN, complex carbohydrates, density estimation, GC-EIMS, cluster analysis

INTRODUCTION

Carbohydrates are one of the three macronutrients in our diet that provide energy [1]. Carbohydrates are carbon-based compounds that contain large quantities of hydroxyl groups. Carbohydrates can be classified into three major groups namely, Monosaccharides (single sugars, e.g. glucose), Oligosaccharides and Polysaccharides. Oligosaccharides are typically chains of up to 20 monosaccharides. Polymers of monosaccharide that contain more than 20 monosaccharides are called Polysaccharides or glycans. These macromolecules have long chains, high molecular weight, very flexible structures, and are characterized by complex chemical structures. The latter two groups of carbohydrates are known as complex carbohydrates because of the number of sugar units in their structure and complexity of linkage. Recent

research in the field of complex carbohydrates has indicated that these macromolecules perform important physiological functions and their role is not just limited to energy metabolism as was initially thought. Complex Carbohydrates are important components of many biological processes such as regulation and mediation of immune response system and organism immunology [2-4]; cell signaling [5, 6, 7]; growth and development [8, 9, 10]; mammalian embryonic development [11] and gene expression [12, 13]. Recent studies have shown that changes in the structure of glycolipid and glycoprotein oligosaccharides in mammals could be tumor-related changes [13]. Complex Carbohydrates such as N-linked oligosaccharides have also been shown to correlate with certain inherited and communicable diseases [14, 15]. The body of research that signifies the importance of complex carbohydrates is rapidly growing. This has led to extensive research effort that studies complex carbohydrates with a variety of goals in mind (e.g. towards the development of novel therapeutic, diagnostics, drug strategies and nanomedicine [15, 16, 17, 18, 19]). The structures of complex carbohydrates can be complicated and extensive. These complex structures are defined by identifying the monomers forming the polymer and their linkage positions. As compared to nucleic acids and protein structures, the structures of complex carbohydrates can be highly variable based on the stereo-chemistry of the base monomers and the linkages between them. A few N-linked oligosaccharides have been shown to have 10,000 different theoretical structures attributed to the extensive stereo-chemistry of these macromolecules. [17] This makes the elucidation of complex carbohydrates based on their structure, a highly rigorous, repetitive and time consuming effort. This identification task is further complicated by the fact that the structure of the same carbohydrate molecule (e.g. glycoprotein) can vary depending upon its source.

One way of obtaining the composition of these macromolecules is by analyzing the GC-EIMS spectra of the PMAAs obtained as a result of methylation of the molecule. Glycosyl linkage analysis of PMAAs using GC-EIMS spectroscopy is an analytical method

commonly used to determine the positions of the linkages between sugar rings in oligosaccharides and polysaccharides [20]. ‘Furthermore, monosaccharides can be distinguished on the basis of their carbonyl group and the number of Carbon atoms’ [21]. Figure 1 depicts how a PMAA is obtained from a monosaccharide.

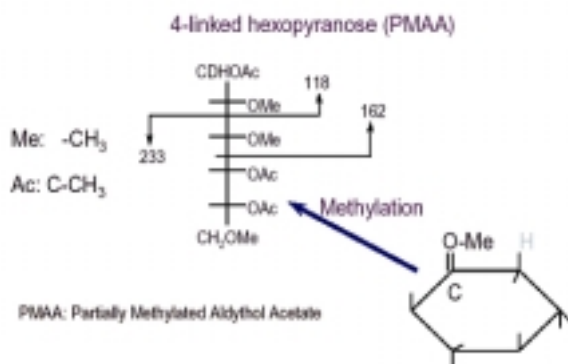


Figure 1. *Methylation of a 4-Linked Hexopyranose producing a PMAA. As indicated, the sugar residue gets separated at a relative density of 118. GC-EIMS produces a spectrum of this PMAA as shown in the fig. 2 below.*

In this paper, we present an approach towards automated structural identification of complex carbohydrates. Here we discuss an approach that combines feature selection (using PCA) and pattern recognition (using k -NN classification) that discovers the natural groupings of the PMAAs. We show that it is possible to determine the type of the PMAA (i.e. mannose, fuctose, arabinose, xylose, galactose, glucose, etc.) with a relatively simple and computationally inexpensive statistical approach. This paper is organized as follows. After Introduction, we offer a discussion of the previously performed relevant work in section 2. In section 3, we discuss our approach to feature selection and the results obtained in order to avoid the ‘curse of dimensionality’. Section 4, discusses our approach to classification, namely, our implementation of the k -NN classification and our best results. Section 4 offers a discussion of the results and offers some conclusive remarks. The final section offers a list of the literatures sighted in this article.

I. PREVIOUS WORK

Previous work done with regards to identification of chemical compounds have shown that pattern recognition techniques could be successfully applied to their classification [22, 23, 24]. Such techniques have been applied to complex carbohydrate data sets and interesting results have been obtained. Techniques like Bayesian Methods, k -Nearest Neighbor, SIMCA, [25, 26, 27, 28] and more extensively (and recently) Neural

Networks [18, 29, 30, 31] have been used for structural identification. Most of these techniques use the Nuclear Magnetic Resonance (NMR) spectra instead of PMAA spectra of these molecules. A comparison of our results and those used in the above references is made in later sections.

II. FEATURE SELECTION WITH PRINCIPAL COMPONENT ANALYSIS

Discrete Karhunen-Loeve expansion, or as it is also known as PCA, is a linear technique that is often used to reduce the dimensionality of a data-set that suffers from the ‘curse of dimensionality’. This method is therefore used as a mechanism to select a subset of features that describe the patterns available in the dataset. PCA produces a set of orthonormal vectors called the feature vectors from the scatter matrix of the dataset. A subset of these *feature vectors* is chosen to represent the data set with an acceptable mean-squared error. [32, 33, 34] Using these feature vectors, one can then produce a feature ‘template’ for each pattern in the dataset. These templates contain *feature values* that can be thought of as linear combination of the original variables of the data. The advantage here is that the feature templates often have much lower dimensionality than the original data and can often be used, in place of the original data, for classification purposes.

Our training and testing sets consisted of GC-EIMS spectra of 11 different types of PMAAs. The training set contained 363 spectra and the testing set consisted of 242 spectra. For our classification purposes, we consider the 11 types of PMAAs as 11 classes. Each of these spectra contains information about the relative densities of the PMAA derivatives produced as a result of fragmentation. [20] An example of one such spectrum is shown in the Figure 2. Although in this example the spectrum runs a range of 40 to over 350 m/z (mass/charge), the spectra used in our study only contain values for the range 50 through 350 m/z .

The 11 different sugars (pyranosyl and furanosyl rings) of the training and testing sets are D-arabinitol (ara), L-galactitol (fuc), D-galactitol (gal), acetylated amine derivate of D-galactitol (galNac), D-glucitol (glc), acetylated amine derivate of D-glucitol (glcNac), D-mannitol (man), acetylated amine derivate of D-mannitol (manNac), L-mannitol (rha), D-ribitol (rib) and D-Xylitol (xyl). The various linkages found in these molecules are T-,2-,3-,4-,6-linked; 2,3-;2,4-;2,6-;3,4-;3,6-;4,6-;2,3,4-;2,3,6-;2,4,6- and 3,4,6- linked chains. Where T indicates a terminal carbon linkage, and the numbers represent a linkage at different carbon atoms numbered between 1 and 6 in the ring.

Our objective for feature selection is to reduce the dimensionality to a minimum number m from the

original 301 dimensions (50 through 350 m/z) without appreciable loss of information. The hope is that a feature space with lower dimensionality will make it easier to build a model that can correctly classify GC-EIMS spectra of PMAAs into their respective classes.

The details of the algorithm are follows. Let Z be a data-matrix consisting of P spectra (363 in our case) of the training data. Further, Let N be the dimensionality of the data (301 in our case). This matrix can be represented in terms of independent vectors as:

$$Z = \sum_{i=1}^N y_i \phi_i = \Phi Y \quad (1)$$

where, $\Phi = [\phi_1, \phi_2, \dots, \phi_N]$ are the feature vectors, and $Y = [y_1, y_2, \dots, y_N]$ is a matrix containing the feature templates representing Z in the feature space.

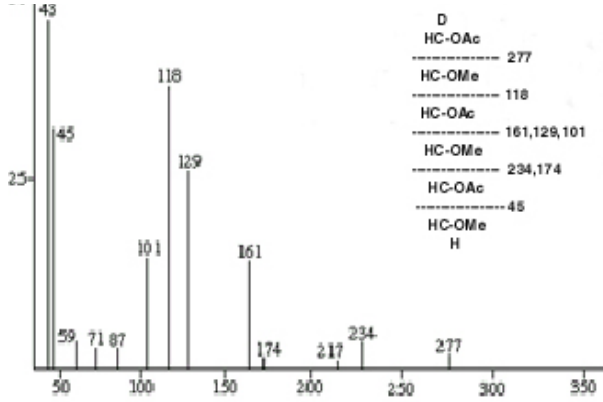


Figure 2. Mass spectrum of PMAA by GC-EIMS of 1,3-linked Hexose. The abscissa represents the mass number (m/z) and the vertical axis indicates the relative density of the PMAA derivatives. The derivatives and their relative density are indicated in the legend.

The ϕ_i 's are orthonormal vectors. The feature templates y_i 's are mutually exclusive (i.e. the covariance matrix of Y is diagonal). Hence,

$$Y = \Phi^T Z \quad (2)$$

Z is $N \times P$ dimensional, with each spectrum residing as a column in the matrix. Our objective is to find m and the set $\hat{\Phi} = [\phi_1, \phi_2, \dots, \phi_m]$ so that $m < n$, and the set $\hat{\Phi}$ can represent Z with the minimum mean squared error.

As a preprocessing step, we first normalized, and then shifted the columns in Z so that they became zero-mean. Equation (2), hence becomes:

$$\hat{Y} = \hat{\Phi}^T (Z - \mu) \quad (3)$$

where, μ is the expected value of columns of Z , and \hat{Y} is a close estimate of the original Y . The next step is to calculate the co-variance matrix of Z .

$$\Sigma_Z = \frac{Z \times Z^T}{N} \quad (4)$$

where, Σ_Z is the covariance matrix of the matrix Z and has dimensionality of $N \times N$. It can be shown [35] the solutions to equation (5) optimize the mean squared error.

$$\Sigma_Z \Phi = \Sigma_Z \lambda \quad (5)$$

Where, λ and columns of Φ are the eigenvalues and corresponding eigenvectors of the scatter matrix. Satisfying equation (5) optimizes the mean squared error. The mean squared error is then given by,

$$\epsilon^2 = \sum_{i=m+1}^N \lambda_i \quad (6)$$

The effectiveness of each feature vector in representing the original data (Z) is determined from its corresponding eigen-value. This is indicated by equation (6). If for instance ϕ_i is dropped from $\hat{\Phi}$, the value of the mean squared error is increased by the corresponding λ_i . The feature vectors are therefore dropped from $\hat{\Phi}$ in the order of increasing values of λ_i starting with ϕ_N .

The above mentioned algorithm was implemented using Matlab. Feature vectors and their corresponding eigen-values were calculated for the spectral data. The importance of each eigen-value was determined using both an eigen-value plotting scheme and by analyzing the percentage contribution of each eigen-value. The eigenvalue plot for GC-EIMS spectra of PMAAs is shown in Figure 3.

The relative percentage contribution of each of the eigenvalues was also analyzed. The relative percentage of the eigenvalues can be found as,

$$f_i = \frac{\lambda_i}{\sum_{i=1}^N \lambda_i} \quad (7)$$

Calculating the percentage contribution using equation (7), we get results as depicted in Figure 3. A 95% threshold was applied to choose the eigenvectors. Starting from the highest eigenvalue, we added the values of the corresponding f_i were added until we reached a sum of 95%. This approach revealed a set $\hat{\Phi}$ that contained only the first 17 (out of 301) feature vectors (since the corresponding eigenvalues added up to 95.5%). We then proceeded to compute the feature templates for each column of Z using the newly computed $\hat{\Phi}$. The feature values were calculated as presented below,

$$\hat{Y}_{m \times P} = \hat{\Phi}_{m \times N}^T (Z - \mu)_{N \times m} \quad (8)$$

where m is the number of selected feature vectors (17 in our experiments). As evident from the equation, we have decreased the dimensionality of the training data set from 301 to m . The columns of the feature matrix \hat{Y} are the feature templates that will be used to classify the data into the 11 classes.

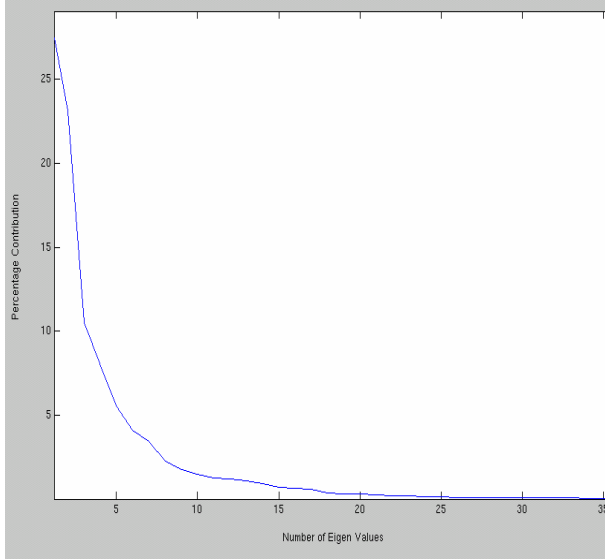


Figure 3. Eigen Values plot to explain the significance of feature vectors. As evident from the figure, eigenvalues greater than 17 can be comfortably dropped since the contribution from the corresponding eigen values is negligibly low.

III. CLASSIFICATION SYSTEM

The features selected from the PMAA spectral data offer a much smaller dataset for classification purposes. For classification of the feature values, initially, *Parzen* density estimation was used in combination with *Bayes'* classification. [35]

Bayes' Classification allows for the calculation of the posteriori probability in terms of a priori probability. Bayes' likelihood classification rule expressed in terms of the a priori probabilities for a two-class case without penalties is given below.

$$f_{H|X}(H_A | x)P_A \geq f_{H|X}(H_B | x)P_B \quad (9)$$

In equation (9), A and B are the two different classes (hypotheses), $f_{H|X}$ is the a priori conditional probability density function given that a value of x was observed, and P_A and P_B are the a priori probabilities for the two hypotheses.

Parzen density estimation, a non-parametric *kernel*-based technique, was used to calculate the conditional probability density functions. Kernel-based density estimation offers the advantage of virtually not requiring any prior knowledge of the density function

that is to be estimated. In this technique, the contribution of every data point to the overall probability distribution function is estimated using what is called a *kernel*. [35] The scaled summation of all the kernels (for all the data points) then provides an estimate for the probability density function.

As expected, the results obtained from one-dimensional models using this approach were not very promising. This was true irrespective of which feature value was chosen as the one dimension of the model. Highly overlapped classes and low classification accuracy was observed in all cases. To improve the results, two-dimensional kernel density estimation was used. For this purpose, the first two feature values (out of the 17 selected) were used for estimation and classification purposes. The 2-dimensional density estimation plot for the chosen feature vectors is shown in Figure 4. As can be observed from the figure, the classes again show highly overlapped regions, as the peaks in the plot represented multiple classes. Because of the persisting poor performance of the algorithm as well as the fact that the complexity of the algorithm increases manifold at high dimensions, we opted to experiment with k -NN classification algorithm [35] for higher dimensional analysis.

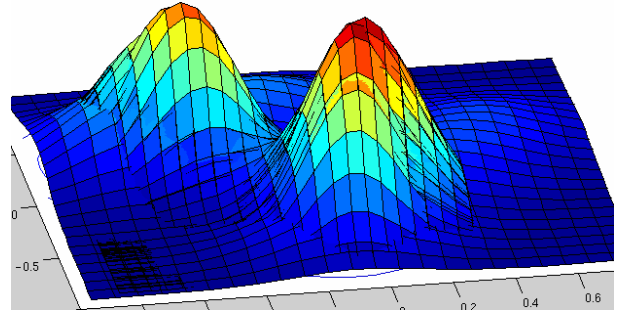


Figure 4. Two-dimensional Kernel density estimation.

k -NN classification is a non-parametric classification technique and has been used for numerous classification studies. In this technique, when new data arrives, the nearest k neighbors (to the new data point) are found and used in a voting scheme in order to determine the class of the new data point (pattern). It can be shown that in the present of sufficient data k -NN classifier is equivalent to Bayes' decision using k -NN density estimation of the posteriori densities. [35] The k -NN classifier is based on the assumption that the classification of a data point is most similar to the classification of other data points that are nearby (in the Euclidian sense). The voting scheme can also take into account any optimized risk functions (if risk factors are provided or can be computed). [36] The algorithm is simple to implement and its error is at best equal to

Bayes' error using k -NN estimation and at most twice Bayes' error k -NN estimation. [35]

k -NN classifier is a computationally simple even in higher dimensional spaces. Further experiments on the feature values were done using k -NN classifiers. The k -NN algorithm for Matlab was adopted from [36].

The k -NN classifier was tested with various values of k for the 17-dimensional feature space. It was found that the lowest error rates were obtained for $k=1$ (both for training and testing). Figure 5 shows the accuracy of the model for different values of k for the test set. The classification model with $k=1$ was chosen and it was tested against the testing data and a success rate of around 23.55% (25.4% for the training set) was obtained. Furthermore, it was observed that the algorithm classified the data into four clusters rather than the original 11 classes.

Table 1, shows the division of the original 11 classes into the 4 clusters that k -NN algorithm has produced. The first cluster contains 7 of the 11 classes, where the second cluster contains 2 classes. The remaining two clusters contain one of the original classes each. As a result, with this model, the classification accuracy for classes 3 and 5 was very high (92.3% and 90.2% for the two classes respectively), while that of the remaining classes was much lower (due to the overlap). To improve the classification accuracy of the remaining classes, we decided to perform the classification for the classes in the first two clusters in several steps. Each of the steps uses the same algorithm as the one described above, except that the subsequent steps only consider a portion of the problem space considered for the previous steps (a "divide and conquer" approach).

In this way, the first step uses the steps described above where classes are clustered into one of the first 4 clusters of classes. For clusters 1 and 2, subsequent steps follow, which then repeat the algorithm for an increasingly smaller subspace until all classes have been separated. The subsequent steps for clusters 1 and 2 are discussed in more detail in the following.

Cluster 1: In step 2 of classification for data that fall into the first cluster, we repeated the entire process, except that we only consider the data for classes 1, 4, 6, 7, 8, 10, and 11 from the training set. This process involved repeating the steps for feature selection as well as those for k -NN classification. The 95% rule, in this case revealed the importance of only the first 9 feature vectors. As a result, the first 9 features values were considered for classification by k -NN. Again, here the case of $k=1$ produced the best results. The overall classification accuracy obtained was 63.77%. This time around, 4 of the 7 data classes contained in the cluster

were correctly classified with accuracy around 90% while the remaining 3 classes still formed one cluster.

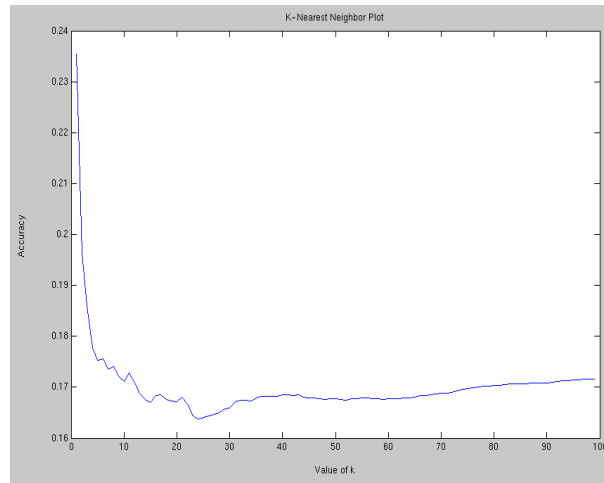


Figure 5. Figure shows that the accuracy of the k NN model is highest at $k=1$ and the accuracy drops with increasing k . The Y-axis represents the accuracy and the X-axis represents the value of k

Table 1: Classes and the molecules associated with each cluster obtained after testing the training model.

Cluster	Classes	Molecules
1	1	D-Arabinitol (Ara)
	4	D-galactitol(galNac)
	6	D-glucitol(glcNac)
	7	D-mannitol(man)
	8	D-mannitol(manNac)
	10	D-ribitol(rib)
	11	D-Xylitol
2	2	L- galactitol(fuc)
	9	L-mannitol(rha)
3	3	D-galactitol(gal)
5	5	D-glucitol(glc)

Further repeat of this cluster analysis did not produce an improvement in the results. Here cluster analysis was performed by taking the training data from these 3 data classes. The process was again repeated. In this case, the 95% rule indicated the importance of only the first 3 feature vectors. The best accuracy attained after the classification was around 54.17% (obtained for $k=2$). The 3 classes in this cluster still showed significant overlap. The three classes were galNac, glcNac, and manNac. A look at the GC-EIMS spectra of these three groups revealed that the spectra of these groups are highly similar. It is also very interesting that the algorithm is able to cluster the spectra into groups that bare structural resemblance.

Cluster 2: The second cluster obtained after the first level of classification contained 2 classes. A reclassification process as explained above was performed on this cluster. The first 11 feature vectors were chosen based on the 95% scoring strategy. The 2 classes in the cluster were classified with an accuracy of 83.33%. This accuracy was obtained at $k=1$. Thus, the re-classification process significantly improved the classification accuracy of the classes of this cluster. Figure 6 demonstrates the overall flowchart of the classification process.

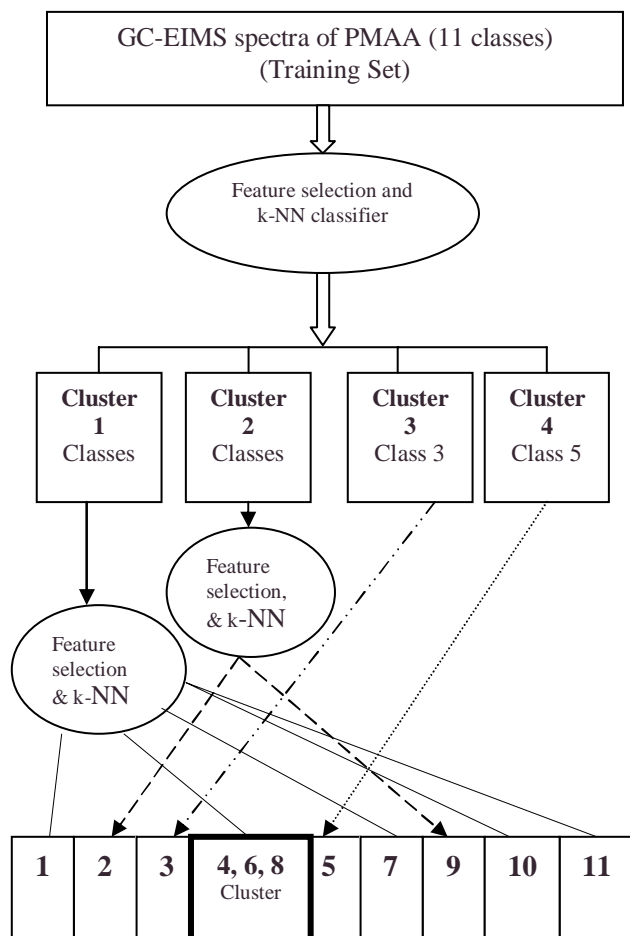


Figure 7. Figure shows the block diagram of the cluster based classification process implemented in this paper.

IV. DISCUSSION AND CONCLUSIVE REMARKS

As discussed above, although our experiments have been preliminary and we have only used linear separation techniques, the training model was able to classify the problem space with high accuracy. A cascading classification system was implemented using a statistical cluster-based approach. Here we have demonstrated that such a system is capable of dividing a problem space into linearly separable subspaces (sub-

problems). The role of PCA was crucial as it helped reduce the dimensionality of the training set from 301 dimensions to at most 17 dimensions. In this problem, the feature space was able to effectively represent the data and helped in simplifying the classification process. 8 of the 11 molecules were classified with a very high accuracy. These molecules are D-arabintol (ara), L-galactitol (fuc), D-galactitol (gal), D-glucitol (glc), D-mannitol (man), L-mannitol (rha), D-ribitol (rib) and D-Xylitol (xyl). k -NN has proved to be an effective method for distinguishing between structurally similar molecules. A similar conclusion was drawn in [28]. The training model has shown poor accuracy for the rest of the 3 classes which are represented by the cluster. A more robust system needs to be adopted for their classification, which may be an extension to the present system.

Previous work in this area has shown that nonlinear techniques such as neural networks can be used to identify the GC-EIMS spectra of PMAAs and NMR spectra of oligosaccharides [18, 31, 32, 33]. Other studies have also shown that simple techniques such as SIMCA, K-NN and PCA can also be applied to obtain substantial structural information, though, most of such studies have focused on identification of the $^1\text{H-NMR}$ spectra rather than mass spectra of PMAAs.

REFERENCES

1. Vozzo, R., Wittert, G., Cocchiario C., Tan, W.C, Mudge, J., Fraser R. and Chapman, I.: *Similar effects of foods high in protein, carbohydrate and fat on subsequent spontaneous food intake in healthy individuals*, Appetite, In Press, Corrected Proof, Available online April, 2003, (<http://www.sciencedirect.com/science/article/B6WB2-48F5GG1-6/2/cfafb1cf082858c157c6a1edb6e0b904>)
2. Hennet, T., Chui, D., Paulson, J.C., and Marth, J.D. (1998) Immune regulation by the ST6Gal sialyltransferase. *Proc. Natl. Acad. Sci. USA* **95**, 4504-4509
3. Wu, A. M., *Complex Carbohydrates in Microbial and Viral infections and vaccine design*, The Molecular Immunology of Complex Carbohydrates-2 **1997** In: Ch. 4
4. Fujimiya, Y., Yamamoto, H., Noji, M. and Suzuki, I. *Peroral effect on tumour progression of soluble β -(1, 6)-glucans prepared by acid treatment from Agaricus blazei. Murr. (Agaricaceae, Higher basidiomycetes)*. International Journal of Medicinal Mushrooms **2** **2000** 43-49.
5. Darvill, A., Bergmann, C., Cervone, F., De Lorenzo, G., Ham, K., Spiro, M. D., York, W. S., and Albersheim, P.: *Oligosaccharins involved in plant growth and host-pathogen interactions*, *Bioch. Soc. Symp* **1995**, 60: 89-94.
6. Sacchettini, J. C., Baum, L. G., Brewer, C. F., 2001. *Multivalent protein-carbohydrate interactions. A new*

- paradigm for super molecular assembly and signal transduction*, *Biochemistry* 40: 3009-3115
7. G. Freshour et al., "Developmental and tissue-specific structural alterations of the cell-wall polysaccharides of *Arabidopsis thaliana* roots," *Plant Physiology*, **1996** 110:1413-29
 8. British Nutrition Foundation's Task Force, **1990**. *Complex Carbohydrates in Foods*, Chapman and Hall Ltd.
 9. Jenkins, D. J.; Kendall, C. W.; Augustin, L. S.; Franceschi, S.; Hamidi, M.; Marchie, A.; Jenkins, A. L., and Axelsen, M., *Glycemic index: overview of implications in health and disease*. *Am J Clin Nutr.* **2002 Jul**; 76(1):266S-73S.
 10. Burkitt, D. P. and Trowell, H. C. *Dietary fibre and western diseases*. *Ir Med J.* **1977 Jun**; 70(9):272-7.
 11. Metzler M, Gertz A, Sarkar M, Schachter H, Schrader JW and Marth JD: *Complex asparagine-linked oligosaccharides are required for morphogenic events during post-implantation development*. *EMBO J* **1994 May**; 13(9):2056-65
 12. Koch, K. E., 1996. *Carbohydrate-modulated gene expression in plants*, *Annu. Rev. Plant. Physiol. Plant Mol. Biol.* **47**: 509-540
 13. Elizabeth F. Hounsell, Mia Young and Michael J. Davies: *Glycoprotein changes in tumors: a renaissance in clinical applications*, *Clinical Sci.* **1997**; 287-93.
 14. Tan J, Dunn J, Jaeken J, Schachter H. *Mutations in the MGAT2 gene controlling complex N-glycan synthesis cause carbohydrate-deficient glycoprotein syndrome type II, an autosomal recessive disease with defective brain development*. *American Journal of Human Genetics*, **1996** 59: 810-817.
 15. Bush C., Martin-Pastor M., Imberty A.: *Structure and Conformation of Complex Carbohydrates of Glycoproteins, Glycolipids and Bacterial Polysaccharides*, *Annu. Rev. Biophys. Biomol. Struct.* **1999**, 28: 269-93.
 16. Koeller, K.M., Wong, C.-H.: *Emerging themes in medicinal glycoscience*. *Nat. Biotechnol.* **2000** 18:835.
 17. Naik RS, Branch OH, Woods AS, Vijaykumar M, Perkins DJ, Nahlen BL, Lal AA, Cotter RJ, Costello CE, Ockenhouse CF, Davidson EA, Gowda DC. : *Glycosylphosphatidylinositol Anchors of Plasmodium falciparum: Molecular Characterization and Naturally Elicited Antibody Response That May Provide Immunity to Malaria Pathogenesis*, **December , 2000** *The Journal of Experimental Medicine*, Volume 192, Number 11, 1563-1576.
 18. Valafar F. and Valafar H.: *CCRC-Net: CCRC-Net: an Internet-based spectral database for complex carbohydrates using artificial neural networks search engines*, *TrAC: Trends Anal. Chem.* **1999**, 18.
 19. Freitas, R.: *Nanomedicine*.
 20. Tadashi T., Katusko Y., Akira K.: *Synthesis and Mass Fragmentographic Analysis of Partially O-Methylated 2-N-methylglucosamines*, *J. Biochem.* **1975** 78: 679-86.
 21. Voet D., Voet J.G., Pratt C.W.: *Fundamentals of Biochemistry*, John Wiley & Sons, Ch : 8.
 22. Tadashi T., Katusko Y., Akira K.: *Synthesis and Mass Fragmentographic Analysis of Partially O-Methylated 2-N-methylglucosamines*, *J. Biochem.* **1975** 78: 679-86.
 23. Song X., Hopke P.K., Bruns M.A., Graham K., Scow K.: *Pattern Recognition of Soil Samples based on the Microbial Fatty Acid Contents*, *Environ. Sci. Technol.*, **1999**, 33(20): 3524-3530.
 24. Arruda, A. F., Goicoechea H.C., Santos M., Campiglia A.D., Olivieri A.C. : *Solid-Liquid Extraction room temperature Phosphorimetry and Pattern Recognition for screening Polycyclic Aromatic Hydrocarbons and Polychlorinated Biphenyls in Water Samples*, *Environ. Sci. Technol.* **2003**, 37(7): 1385-91.
 25. Goux, W.J.: *NMR Pattern Recognition of Peracetylated Mono- and Oligosaccharide Structures. Classification of Residues Using Principal Component Analysis, K-Nearest Neighbor Analysis and SIMCA Class Modeling*, *J. Magn. Res.*, **1989**, 85: 457-469).
 26. Goux W.J., Weber D.S., Okike G., *The Selection of NMR Spectral Features Leading to the Optimum Classification of Residue Types in Peracetylated Oligosaccharide Derivatives*, *J. Magn. Res.*, **1992** 96, 526-540.
 27. Goux W.J., Weber D.S., *The Application of NMR/Pattern Recognition Methods to the Classification of Peracetylated Oligosaccharide Residues: Effects of Intra-class Structure*, *Carbohydr. Res.*, **1992** 233, 65-80.
 28. Anja-Carina S., Adrian G., Christof A., Klaus-Peter N., Hans R. K.: *Use of Global Symmetries in Automated Signal Class Recognition by a Bayesian Method*, *J. Magn. Res.*, **1997**, 165-172.
 29. Meyer B., Hansen T., Nute T., Albersheim P., Darvill A., York W., Sellers J., *Identification of the ¹H-NMR spectra of complex oligosaccharides with artificial neural networks*, **1991**, *Science*, 542-544.
 30. Sellers J., York Y., Albersheim P., Darvill A., Meyer B.: *Identification of the mass spectra of partially methylated alditol acetates by artificial neural networks*. *Carbohydr. Res.*, **1990**, 207: C1-C5.
 31. Radomski J.P., Halbeek H.V., Meyer B.: *Neural network-based recognition of oligosaccharide ¹H-NMR spectra*, **1994**, *Nature Structural Biology* 1: 217-218.
 32. Mallat, S. G.: *A Wavelet Tour of Signal Processing*, 2nd edition **1999**, Ch. 9, Academic Press, San Diego.
 33. Anderson. T. W.: *Introduction to Multivariate Statistical Analysis*, 2nd edition **1984**, Wiley, NY
 34. Lu W.: *Face Recognition by Karhunen Loeve Expansion*, **December 1997**.
 35. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, 2nd edition, **1990**, Ch 9: 401-405, Academic Press 1990.
 36. L. M. Zouhal and T. Denoeux: *An evidence-theoretic k-NN rule with parameter optimization*. *IEEE transactions on Systems, Man and Cybernetics*, **1998**, Part C, 28(2): 263-271.