Computer-Assisted Interpretation of Mass Spectra

Carleton S. Hayek, Fernando J. Pineda, Otis W. Doss III, and Jeffrey S. Lin

Under Defense Advanced Research Projects Agency sponsorship, APL is developing a miniature time-of-flight (TOF) mass spectrometer for early warning against exposure to chemical/biological agents. Intended for operation by a wide range of military and civilian personnel, the instrument must be able to detect and identify pathological agents within minutes. Key to this mission is the spectrometer operator's interpretation of the data. Typically, interpretation of mass spectra has been the realm of professional chemists and biochemists. Other operators must rely on computer classification of the TOF mass spectrometer's output. We describe algorithms that can be used to interpret mass spectra and that have been successful on a limited data set. These algorithms handle precisely known, and partially unknown, signatures. For precisely known signatures, a vector space problem can be formulated to estimate the optimum approximation of the measured spectrum with a combination of stored library signatures of threat agents. For partially unknown signatures, a Bayesian probabilistic approach has been taken to relate the potentially variable signature of a bacterial threat to likelihoods of chemical composition of bacterial lipids. (Keywords: Computer classification, MALDI, Mass spectrum.)

INTRODUCTION

The goal of the APL Miniature Time-of-Flight (TOF) Mass Spectrometer Program is to produce a field-portable device (see McLoughlin et al., this issue). This instrument is intended to be used by military or emergency civilian personnel to detect the presence of chemical and biological warfare (CBW) agents while there is still time to minimize their effects. The ability to rapidly detect and classify chemical or biological threats is critical to the safety and effectiveness of military forces and civilian populations.

Key to this mission is computer-assisted interpretation of the large quantity of data produced by the mass spectrometer. The complex output must be made accessible to operators who have no background in mass spectrometry. The software being designed to accomplish this part of the mission is designated the Threat Identification System (TIDS). Challenges to the development of an automated threat identifier are:

- Possibility of highly noisy and cluttered background

- Complexity of agent molecules (hence, complexity of mass spectral signatures), combined with similarity of their basic bio-organic chemistry
- Variability of bacterial molecular signatures as a result of growth conditions and incompletely understood ionization physics
- Need for the identification algorithm to perform at high probability of detection with low probability of false alarm, while being able to analyze low concentrations of agents or agent mixtures

The TOF mass spectrometer may be deployed when the exact signature masses of the pathological agents are well characterized (e.g., from intelligence collection of weaponized agents) or when they are only approximately characterized. With detailed prior knowledge, the agent can be identified by pattern matching to a preconstructed library. With only approximate prior knowledge, a broader approach may be required; for example, the processor may apply a set of bio-organic molecular consistency rules for agent identification. We report here on progress toward accomplishing the goal of computer-assisted spectrum interpretation.

MASS SPECTROMETER OPERATION AND SIGNATURES OF SELECTED ORGANIC COMPOUNDS

Figure 1 shows a simplified diagram of a TOF mass spectrometer. The sample is introduced in the inlet, which is under vacuum. Matrix-assisted laser desorption/ionization (MALDI) is used to ionize the sample, typically resulting in a net single charge on the molecule. Ionized sample molecules, or ionized molecule fragments, are then accelerated by a potential difference into the analyzer section. Because the fragments have equal kinetic energies, their velocities differ in proportion to the square root of their mass ratios. Fragments traveling at their respective speeds will arrive at different times at the detector. Voltage output from the detector will increase with increasing number of fragments arriving simultaneously, thereby indicating the abundance of fragments of each mass.[1] (Fragments of the same mass but multiply charged will not arrive simultaneously; this, however, can be accounted for in the abundance calculation.)

Example spectra of CBW agents are shown in Fig. 2. Note the characteristic pattern of the relative intensities along the mass axis for each agent. For testing purposes, simulants for CBW agents are used as samples. Example spectra for simulants prepared at APL are shown in Fig. 3. Ongoing signature characterization studies will determine the variability of these signatures as a function of growth and environmental conditions, sample preparation, and mass spectrometer configuration and operating parameters.

METHODS FOR AGENT IDENTIFICATION

We have designed two mass spectrum signal processing algorithms, corresponding to the two identification scenarios mentioned earlier. One is a multivariate linear least-squares regression of the unknown spectrum to a spectra library, and the second is a belief network capable of classifying organic substances on the basis of their chemical (i.e., phospholipid) content.

Multivariate Least-Squares Regression

In the first identification scenario, the field operator has measured a sample and must determine whether the mass spectrum matches one threat signature or a combination of threat signatures stored in a library.

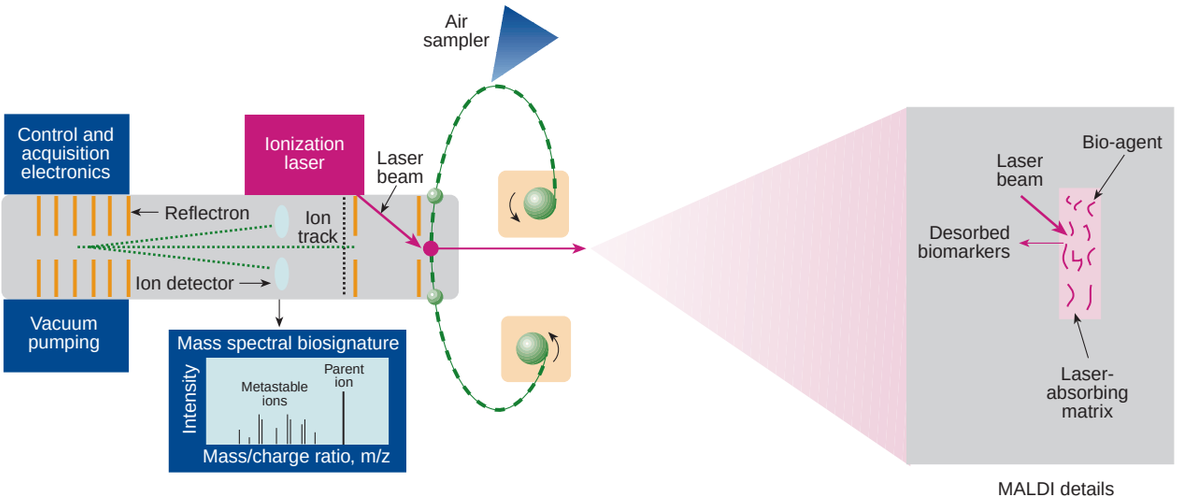


Figure 1. Miniature time-of-flight mass spectrometer biodetector system (MALDI = matrix-assisted laser desorption/ionization).

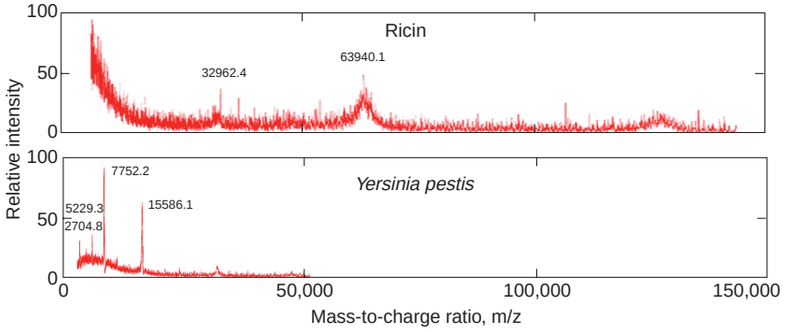


Figure 2. Preliminary mass spectrum signatures of ricin (toxin made from castor beans) and *Yersinia pestis* (gram-negative organism that causes plague).

The mass spectrum can be specified by a multirow, two-column table. The first column is "m/z" (mass divided by charge, where charge is typically 1). The second column is "intensity," i.e., the magnitude of the voltage induced by the ions on the multichannel ion detector plate in the TOF mass spectrometer. When the mass value sequence is the same for a group of spectra, they may be represented by their intensity columns alone, i.e., they can be represented as "intensity vectors." In this section we assume that the exact masses of the predominant fragments of the threat agents are known, unique, and stored in a library as intensity vectors.

Previous work on this problem has demonstrated that mass spectra of mixtures of substances (mixtures of "analytes") result in combinations of their respective mass spectra. Following the results of Platt et al.,[2] we cast the threat identification problem as a vector space optimization problem.

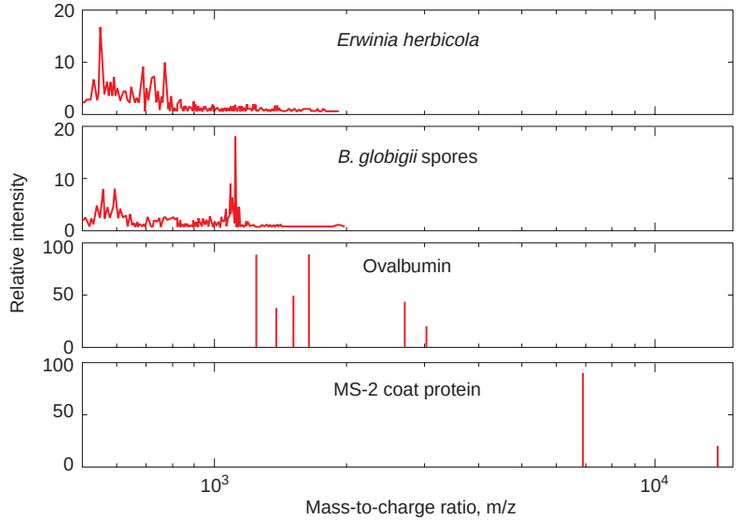


Figure 3. Preliminary mass spectrum signatures of four threat agent simulants used for testing purposes: *Erwinia herbicola*, a simulant for a vegetative bacterium such as plague; *Bacillus globigii*, a spore-forming bacterium similar to anthrax; ovalbumin, a chicken egg protein developed as a simulant for toxins such as botulinium; and MS-2, a simulant for pathogenic viruses. The bottom two panels show peaks only.

In the noiseless case, the problem is to determine optimum weighting coefficients β_k such that the weighted sum of the intensity vectors in the library best match the unknown spectrum. (The weighted sum will not exactly match the unknown spectrum because of slight differences in the absolute intensities of the library signature relative to a newly collected signature.) The approximation is written as

$$\beta_1 \mathbf{L}_1 + \beta_2 \mathbf{L}_2 + \ldots \beta_N \mathbf{L}_N \approx \mathbf{U}, \qquad (1)$$

where $\mathbf{L}_k$ is the abundance vector for library element k, N is the number of elements (i.e., threat agents) in the library, and $\mathbf{U}$ is the abundance vector for the unknown. Non-zero values of β_k indicate that at least one signature line of library member $\mathbf{L}_k$ is present in the sample, assuming a noiseless, interference-less background.

Actual mass spectra intensity vectors are the result of noisy measurements, with possible interference from other substances in the environment. Noise impacts the library elements, the unknown spectrum, and the approach to identifying the optimum mixture coefficients $\boldsymbol{\beta}$.

Assuming that in Eq. 1 the model for the relationship between unknown and library elements is $\mathbf{U} = \mathbf{L}\beta + \boldsymbol{\xi}$ (with $\boldsymbol{\xi}$ being the noise in the measurement of $\mathbf{U}$), then the solution for the minimum variance, unbiased estimator $\boldsymbol{\beta}$ is given by the Gauss-Markov theorem[3] as

$$\boldsymbol{\beta} = (\mathbf{L}^T \mathbf{K}^{-1} \mathbf{L})^{-1} \mathbf{L}^T \mathbf{K}^{-1} \mathbf{U},$$

where T means transpose and $\mathbf{K}^{-1}$ is the covariance matrix of $\boldsymbol{\xi}$.

The covariance matrix of $\boldsymbol{\beta}$ is given by

$$\mathrm{cov}(\boldsymbol{\beta}) = (\mathbf{L}^T \mathbf{K}^{-1} \mathbf{L})^{-1}.$$

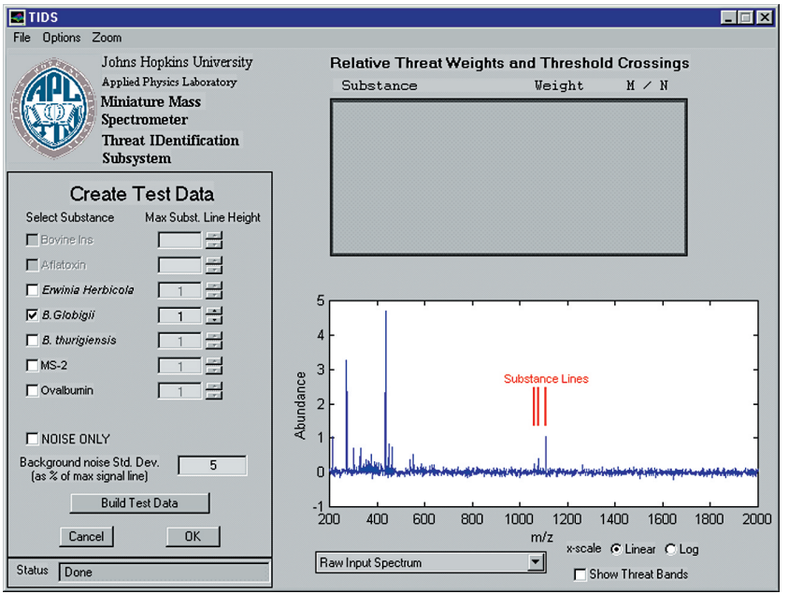
The diagonal elements of $\mathrm{cov}(\boldsymbol{\beta})$ are the variances of the individual components β_k, and indicate the uncertainty in the estimated weights of each library vector caused by the measurement noise. The weights β for the various

library elements, together with the number of signature lines in each library element that appear in the unknown (relative to total number of lines in the respective library element), are presented to the operator.

These computations for TOF mass spectrometer multivariate regression were implemented in TIDS Version 1 software. The software has been run on a 166-MHz PC desktop and a PC laptop, both running Windows 95. Preliminary tests with the available data (the four simulants *Bacillus globigii*, *Erwinia herbicola*, MS-2, and ovalbumin in a noiseless, interference-less environment) showed that each simulant could be readily identified.

As an example, Fig. 4a shows an "unknown" mass spectrum (*B. globigii* spectrum taken with a commercial mass spectrometer, the Kratos MALDI 3) that is to be classified by the processor. Previously stored in the processor's "threat" library were signature lines from seven substances derived from a training set of spectra collected on an earlier date. Three lines in the unknown spectrum correspond to the previously identified *B. globigii* signature and are clearly visible in the range around 1100 m/z. Figure 4b shows the detected lines overlaid on the thresholded spectrum. The probability of false alarms (P_{fa}) for the threshold process was set to 10^{-4}, or 1 in 10,000. Also shown in Fig. 4b, within the box labeled "Relative Threat Weights and Threshold Crossings," are the results of processing the detected peaks with the identification algorithms. As the figure illustrates, all three substance lines were detected and correctly classified as belonging to *B. globigii*.

(a)



(b)

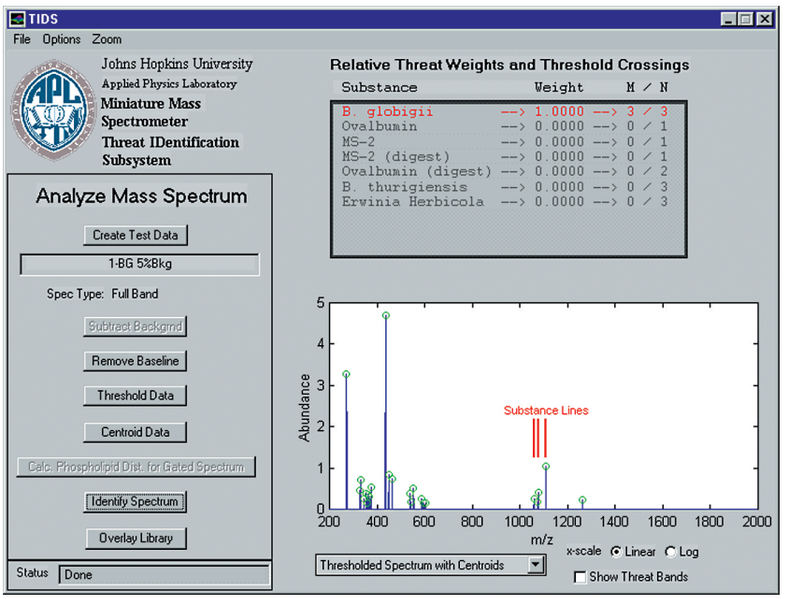


Figure 4. (a) Mass spectrum of an "unknown" (actually the simulant *B. globigii*) presented to the operator by the APL TOF mass spectrometer TIDS after data collection and prior to signature classification processing. The red "Substance Lines" shown here are for the reader's benefit and would not be available to the operator. (b) Correct classification of the unknown in Fig. 4a by multivariate regression to a library of stored mass spectra signatures derived from a separate training data set.

Refinement of this approach will require additional data to characterize the signature intensity and noise distributions. We can then invoke a match confidence measure we have derived[4] to inform the operator of the likelihood that the β_k values presented are a result of a true match of the library with the unknown or of coincidence given the noise environment.

Probabilistic Graphical Modeling for Phospholipid Identification

In the second identification scenario, the field operator has measured a sample and must determine whether the mass spectrum just measured matches one threat agent or a combination of threat agents based on

models of the chemical composition of the threat agent. To demonstrate graphical modeling techniques, we are developing models for phospholipid analysis of MALDI-TOF spectra. Membrane lipids are important biomarkers that have been used to classify bacterial species. In particular, desorption mass spectra obtained from lysed bacteria have been used to distinguish both Gram's stain and species.[5,6]

The most common phospholipids are composed of a glycerol phosphate core, two fatty acids, and a polar head group (Fig. 5). The composition of the polar head group determines the *phospholipid class*. Classifying bacterial species by phospholipid analysis is challenging because the phospholipid content of a given species can be quite variable. For example, the distribution of fatty acids in a given species depends on factors such as culture temperature and the growth phase during which the culture is harvested. The distribution of polar head groups is much less sensitive to growth conditions and has been used by itself to differentiate species.[7,8] Nevertheless, it is far from clear whether phospholipid content alone constitutes sufficient statistics for bacterial classification. Thus, we are constraining our efforts to the characterization of phospholipids in mass spectra, and we defer to a later time the question of whether phospholipid characterization provides sufficient statistics for bacterial classification.

Our approach is based on probabilistic graphical models. This approach is well established in application areas with difficult data-analysis tasks such as speech processing and multisensor fusion. Some well-known examples of graphical models are hidden Markov models, influence diagrams, and Bayesian belief networks.[9] To illustrate the approach, consider Fig. 6.

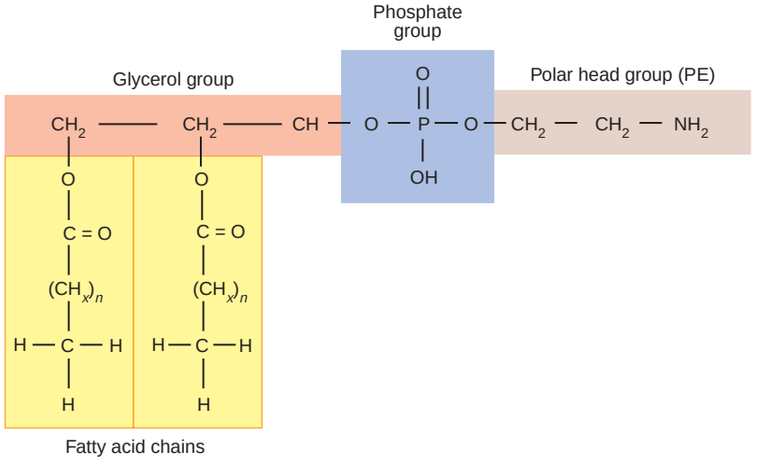


Figure 5. A typical phospholipid consists of four groups. The glycerol group and the phosphate group are always the same. The variable parts are the fatty acid group and the polar head group. The particular phospholipid shown is PE C(34:1). This notation indicates that the polar head group is phosphatidyl-ethanolamine (PE) and that there are 34 carbons and 1 double bond in the fatty acid chains.

This graph represents a joint distribution of random variables as the product of conditional distributions. Each vertex is associated with a conditional distribution wherein the random variable associated with each node is conditioned on the random variables leading to it via directed arcs. In particular, the simple four-layer graph in Fig. 6 represents the joint distribution function

$$P(lipid, H, E),$$

where *lipid* is the random variable that represents the presence or absence of the phospholipid in the spectrum, E is the set of instantiated variables (variables at the bottom the graph), and H is the set of so-called *hidden* variables (all other variables). The variables in the topmost layer are hidden variables that capture prior knowledge concerning the likelihood that a given polar head group (say, PE) and a given fatty acid group (say, C34:1) are found in the sample. The topmost layer is where we account for expert knowledge about growth conditions and polar head group distributions. The second layer contains the random variable *lipid*. This is the query variable whose posterior conditional probability, $P(lipid \mid E)$, we ultimately seek to evaluate. The variables in the third layer are hidden variables that represent the detectable species that could be formed from the fragments and adducts of the phospholipid. Figure 6 models simple biochemical and detection processes. In particular, it models just one fragmentation pathway (wherein the polar head group detaches from the fatty acid group) and three adduct pathways (wherein a sodium, potassium, or calcium ion bonds to the polar head group). In a more realistic model, one would have to account for all the likely fragments and adducts. The variables in the bottommost layer are the evidence variables, which represent features extracted from a mass spectrum.

There are two classes of features in our simple model. The first are "detection" features, which depend on one detectable species each. They are set to the value "true" if a line is detected in the spectrum at the predicted mass. If no line is found, the random variable is set to the value "false." Second are "relative intensity" features, which depend on two detectable species and are set to the value "true" if the two amplitudes of the detected lines are within acceptable ranges of each other. If the relative amplitudes are outside the acceptable ranges, the corresponding relative intensity feature is set to the value "false." If

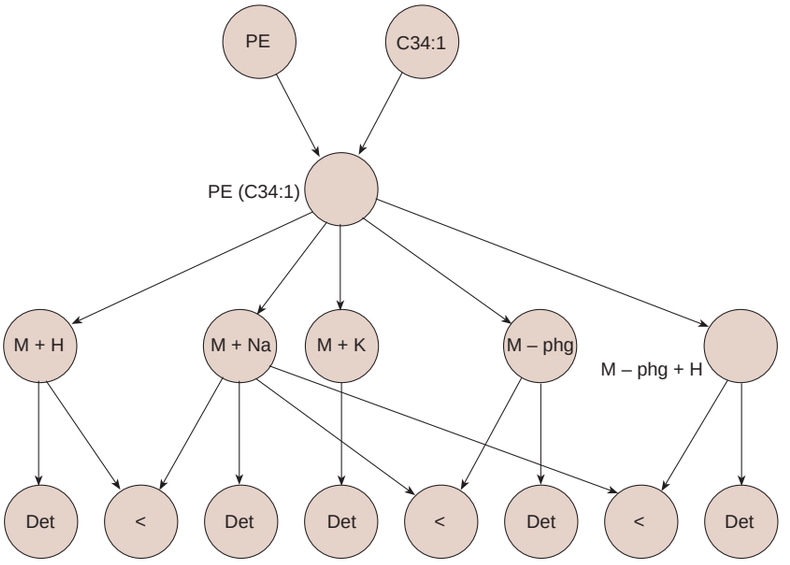


Figure 6. A graphical model representing the joint distribution associated with the phospholipid in Fig. 5, PE (C34:1). The joint distribution depends on 16 binary random variables. The topmost two variables correspond to the prior probabilities associated with the occurrence of the PE polar head group and the C34:1 fatty acid chain. The single node in the second layer is the one whose posterior probability we wish to calculate. The five nodes in the third layer are used to represent the conditional probabilities of forming hydrogen (M + H), sodium (M + Na), or potassium (M + K) adducts, or of losing the polar head group (M − phg), or losing the polar head group and picking up a hydrogen (M − phg + H). The bottom layer consists of variables that represent the events of detecting lines at the appropriate masses (Det) and variables that represent amplitude relationships between detected lines (<).

only one of the detectable species is detected, a relative intensity feature cannot be calculated and, consequently, the corresponding random variable cannot be set to any value and must be treated as a hidden variable. Hidden variables are the mechanism used by probabilistic models for dealing with missing data.

Once the belief network is defined, we can use it to solve the fundamental problem of data analysis, which is to determine the likelihood of a conclusion given the evidence. In this case, the conclusion is whether the phospholipid in question is present (*lipid = true*) or absent (*lipid = false*). Bayes' rule tells us to evaluate $P(lipid \mid E)$ in the following way:

$$P(lipid \mid E) = \frac{\sum_{H} P(lipid, H, E)}{\sum_{H, lipid} P(lipid, H, E)}. \qquad (2)$$

Figure 7 shows the architecture of a prototype system we have developed to perform these calculations. We use the bucket-elimination variant of the *junction tree* algorithm,[10] which is related to the forward-backward algorithm commonly used in hidden Markov models. Figure 8 shows the interface presented to the user by the prototype system. The data in the file selected by the user are plotted in the window at the upper right. The posterior probability $P(lipid \mid E)$ is computed for several hundred phosopholipids and displayed in order of decreasing probability, in the lower right-hand window. A trained analyst may step through each of the candidate phospholipids and examine the detailed evidence used to calculate the corresponding posterior probability. A graphical representation of some of the evidence is overlaid on the displayed spectrum in Fig. 8.

The system we have developed is a prototype intended to explore the feasibility of applying graphical models to phospholipid analysis. We have completed the framework for the system and are currently improving both the preprocessing algorithms and the graphical model. To do the latter we are working with chemists to characterize the expected pathways for fragmentation and adduct formation. Initial results suggest that probabilistic calculations based on graphical methods are tractable and robust. As our understanding of the chemistry and measurement processes becomes more sophisticated, it is inevitable that our graphical models will become more complex. It is likely that the models will prove intractable for the junction tree algorithm. To handle intractable graphs, APL has used internal independent research and development resources to develop a powerful algorithm for approximate inference on highly connected graphs.[11]

SUMMARY

The goal of computer-assisted identification of mass spectra has been approached in the relatively low mass chemical compound arena as demonstrated by commercially available programs such as the National Institute for Standards and Technology (NIST) Mass Spectrometer Search Program[12] and the Probability-Based Matching Program.[13] However, the high masses and complexity of CBW agent molecules relevant to the APL TOF Mass Spectrometer Program, the different ionization technique (MALDI) used in the instrument to emphasize the molecules' unique signatures, and the lack of the equivalent of the NIST/Wiley libraries require new approaches. Working with colleagues at APL, the Army Medical Research Institute of Infectious Diseases, the University of Maryland, College Park, and The Johns Hopkins University School

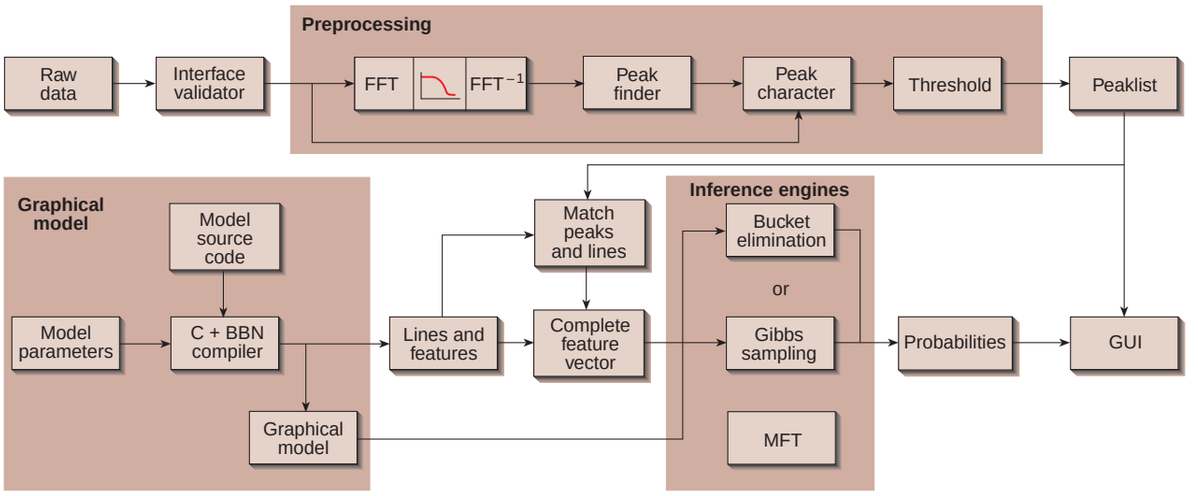


Figure 7. Architecture of "pTool," the application that demonstrates the concepts described in the text. pTool has three major components. The first is the preprocessing component, which extracts a list of peaks from a raw spectrum. The second is the graphical modeling component. Graphical models are described in a special-purpose modeling language that is implemented as an embedded language in "C." Finally, graphical models are evaluated by the inference engines, which use either exact (bucket elimination) or approximate (Gibbs sampling) algorithms. We will shortly implement an approximate algorithm based on mean field theory (MFT). FFT = fast Fourier transform; GUI = graphical user interface.

of Medicine, we have developed two techniques for handling either consistent or variable toxin and CBW agent signatures. As more findings regarding the agents' signatures become available, their implications for computer-assisted spectrum identification will be factored into algorithm development.

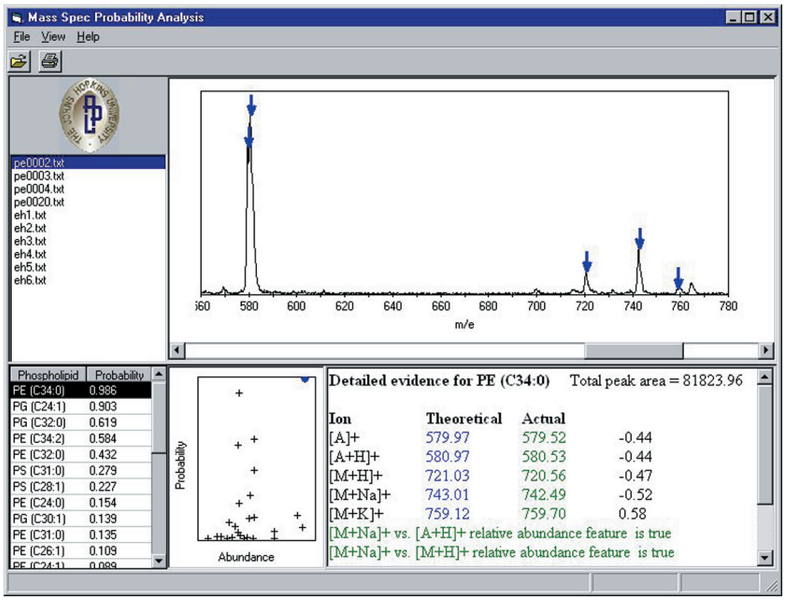


Figure 8. This screen capture shows the graphical user interface presented to the user. The box at the top left allows the user to select a data file containing a spectrum (displayed at the top right). Arrows mark the predicted positions of lines corresponding to specific phospholipids. In this case, the five lines correspond to the phospholipid PE (C34:1). The list of phospholipids at the lower left also gives the posterior probability that a given phospholipid is actually present in the spectrum. The scatter plot at the bottom has one point per phospholipid and provides a quick look for comparing the posterior probabilities with a heuristic measure of relative abundance.

REFERENCES

[1] Bryden, W. A., Benson, R. C., Ecelberger, S. A., Phillips, T. E., Cotter, R. J., and Fenselau, C., "The Tiny-TOF Mass Spectrometer for Chemical and Biological Testing," *Johns Hopkins APL Tech. Dig.* **16**, 296–310 (1995).

[2] Platt, J. A., Uy, O. M., Heller, D. N., Cotter, R. J., and Fenselau, C., "Computer-Based Linear Regression Analysis of Desorption Mass Spectra of Microorganisms," *Anal. Chem.* **60**, 1415–1419 (1988).

[3] Stark, H., and Woods, J., *Probability, Random Processes, and Estimation Theory for Engineers*, Prentice Hall, Englewood Cliffs, NJ (1994).

[4] Hayek, C. S., "Identification of Unknown Mass Spectra Using Statistical Estimation," STX-97-097, JHU/APL, Laurel, MD (13 Jun 1997).

[5] Heller, D. N., Fenselau, C., Cotter, R. J., Demirev, P., Olthoff, J. K., et al., "Mass-Spectral Analysis of Complex Lipids Desorbed Directly from Lyophilized Membranes and Cells," *Biochem. Biophys. Res. Commun.* **142**, 194–199 (1987).

[6] Heller, D. N., Cotter, R. J., Fenselau, C., and Uy, O. M., *Anal. Chem.* **59**, 2806–2809 (1987).

[7] Lechevalier, M. P., "Lipids in Bacterial Taxonomy—Taxonomists View," *CRC Crit. Rev. Microbiol.* **5**, 109–210 (1977).

[8] Goldfine, H., "Lipids of Prokaryotes—Structure and Distribution," in *Membrane Lipids of Prokaryotes*, S. Razin and S. Rotern (eds.), Academic Press, New York, pp. 1–43 (1982).

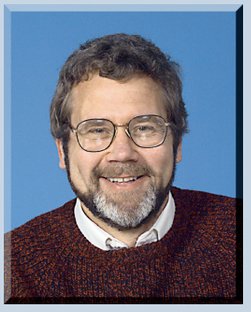
[9] Jordan, M. I., *Learning in Graphical Models*, MIT Press, Cambridge, MA (1999).

[10] Lauritzen, S. L., and Spiegelhalter, D. J., "Local Computations with Probabilities on Graphical Structures and Their Applications to Expert Systems," *J. Roy. Statist. Soc. Ser. B* **50**, 157–224 (1988).

[11]Pineda, F. J., and Wang, I.-J, "Phase-Space Field Theory for Sigmoid Belief Networks," RSI-99-003, JHU/APL, Laurel, MD (1999).
[12]Stein, S. E., "Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification," *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).
[13]McLafferty, F. W., Zhang, M. Y., Stauffer, D. B., and Loh, S. Y., "Comparison of Algorithms and Databases for Matching Unknown Mass Spectra," *J. Am. Soc. Mass Spectrom.* **9**, 92–95 (1998).

ACKNOWLEDGMENTS: We would like to thank Mildred Donlon of DARPA for the financial support to carry out this work, Harvey Ko and Wayne Bryden of APL for helpful engineering and biochemistry discussions, and Catherine Fenselau at the University of Maryland, College Park, and Robert Cotter of The Johns Hopkins School of Medicine for key criticisms and observations.

THE AUTHORS



C. SCOTT HAYEK is a member of APL's Principal Professional Staff. He received B.S. and M.S. degrees in physics from the University of Maryland. He joined APL's Submarine Technology Department in 1978, working in ocean physics and signal processing, and also worked for 3 years in the Naval Warfare Analysis Department in radar signal intelligence. Mr. Hayek has been the principal investigator and test scientist for various sea tests studying the performance limits of passive and active sonar. His current work is split between ocean acoustic surveillance and counterproliferaton R&D. He is a member of the Acoustical Society of America. His e-mail address is scott.hayek@jhuapl.edu.

FERNANDO J. PINEDA is on the APL Principal Professional Staff and is a member of the System and Information Sciences Group of the Research and Technology Development Center (RTDC). He earned B.S. and Ph.D. degrees in physics from the Massachusetts Institute of Technology and the University of Maryland, College Park, respectively. He serves as Program Manager for the Applied Mathematics Program in the RTDC and is a lecturer in the Computer Science Department of The Johns Hopkins University. Dr. Pineda has research interests in neural computation, machine learning, analog VLSI, statistical physics, and quantum physics. He has served on the editorial boards of *Neural Computation*, *Applied Intelligence*, *Neural Networks*, *IEEE Transactions on Neural Networks*, and the *Johns Hopkins APL Technical Digest*. His e-mail address is fernando.pineda@jhuapl.edu.

OTIS W. DOSS III received his B.A. in mathematics from the University of Virginia in 1980 and his M.S. in numerical science from The Johns Hopkins University G. W. C. Whiting School of Engineering in 1984. Mr. Doss has been working at APL as a resident subcontractor since 1993 and currently is a member of the Signal and Information Processing Group. Since coming to APL, he has been involved in developing signal processing algorithms and software for both active and passive underwater acoustic applications. His current interests involve developing graphical user interfaces for tactical display of passive underwater acoustic data. Mr. Doss is a member of Phi Beta Kappa. His e-mail address is otis.doss@jhuapl.edu.

JEFFREY S. LIN received a B.S.E. degree in mechanical/aerospace engineering from Princeton University in 1986 and an M.S. degree in computer science from The Johns Hopkins University in 1989. He is currently working on a doctorate in materials science and engineering at JHU. Mr. Lin is a member of the System and Information Sciences Group of the Research and Technology Development Center, and develops systems for automated machinery diagnostics and nondestructive evaluation of materials. His e-mail address is jeffrey.lin@jhuapl.edu.