Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Ant Colony Optimization for Biomarker Identification from MALDI-TOF Mass Spectra

Habtom W. Ressom, Rency S. Varghese, Eduard Orvisky, Steven K. Drake, Glen L. Hortin, Mohamed Abdel-Hamid, Christopher A. Loffredo, and Radoslav Goldman

Abstract—**We present a novel method that combines ant colony optimization with support vector machines (ACO-SVM) to select candidate biomarkers from MALDI-TOF serum profiles of hepatocellular carcinoma (HCC) patients and matched controls. The method identified relevant mass points that achieve high sensitivity and specificity in distinguishing HCC patients from healthy individuals. The results indicate that the MALDI-TOF technology could provide the means to discover novel biomarkers for HCC.**

I. INTRODUCTION

Analysis of peptides by MALDI-TOF mass spectrometry (MS) is an emerging technology for biomarker discovery. The method has a great potential to identify a panel of biomarkers relevant for early diagnosis of complex diseases such as cancer. Several laboratories have demonstrated the feasibility of selecting peptides in MALDI-TOF spectra for disease classification [1-4].

In our previous work [2, 5], we introduced a computational method that combines particle swarm optimization (PSO) with support vector machines (SVMs) for optimal selection of m/z values from SELDI-QqTOF and MALDI-TOF spectra. A limitation of the PSO algorithm is that it is not tailored for discrete optimization. We used PSO to search for discrete locations in high dimensional space by rounding the positions of the particles to the closest discrete location. In this paper, we present an alternative swarm intelligence-based approach known as ant colony optimization (ACO) that is particularly suitable for discrete optimization. We combined ACO with SVMs to identify the most relevant features (mass points). The algorithm lists these features in the order of their significance in predicting disease state. This will help prioritize candidate protein markers and panels for validation, which leads to assay development applicable to clinical settings.

The paper is organized as follows. Section II introduces our proposed ACO-SVM algorithm. Section III presents samples used in this study, sample preparation methods used to generate mass spectra, data preprocessing methods applied, and biomarkers identified by the ACO-SVM algorithm. Section IV concludes the paper.

II. ACO-SVM

Defined by [6], ACO studies artificial systems that take inspiration from the behavior of real ant colonies. The basic idea of ACO is that a large number of simple artificial agents are able to build good solutions to solve hard combinatorial optimization problems via low-level based communications. Real ants cooperate in their search for food by depositing chemical traces (pheromones) on the ground. Artificial ants cooperate by using a common memory that corresponds to the pheromone deposited by real ants. The artificial pheromone is accumulated at run-time through a learning mechanism. Artificial ants are implemented as parallel processes whose role is to build problem solutions using a constructive procedure driven by a combination of artificial pheromone and a heuristic function to evaluate successive constructive steps.

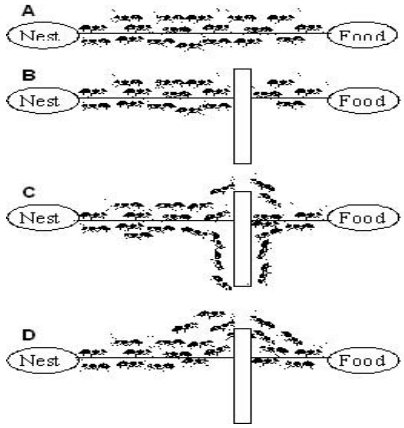


Fig. 1. Pheromone trail following of ants.

Figure 1 illustrates how ants select the shortest trail to fetch their food. The top figure (A) shows a single trail that all ants use to bring food to their nest. An obstacle is placed preventing ants to directly access the food (B). Initially, there is an equal chance for ants to take one of the two trails (note that the upper trail is shorter than the lower trail) (C). Later, ants choose to take the shorter trail (D) as those who used this trail come back to the nest faster than the others that use the second trail. As a result, more and more

This work was supported in part by U.S. Army Medical Research and Material Command, Prostate Cancer Research Program grant DAMD17-02-1-0057 and American Cancer Society grant CRTG-02-245-01-CCE awarded to RG.

H. W. Ressom, R. S. Varghese, E. Orvisky, C. A. Loffredo, and R. Goldman are with Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC (e-mail: hwr@georgetown.edu)

S. K. Drake and G. L. Hortin are with the Clinical Chemistry Service, Department of Laboratory Medicine, NIH, Bethesda, MD.

M.M. Abdel-Hamid is with the Viral Hepatitis Research Laboratory, NHTMRI, Cairo, Egypt.

pheromones will be deposited in the shorter trail over time thereby attracting the ants to use this trail.

We propose to use ACO for feature selection. To accomplish this, we use the probability function below:

$$P_i(t) = \frac{(\tau_i(t))^{\alpha} \eta_i^{\beta}}{\sum_i (\tau_i(t))^{\alpha} \eta_i^{\beta}}$$

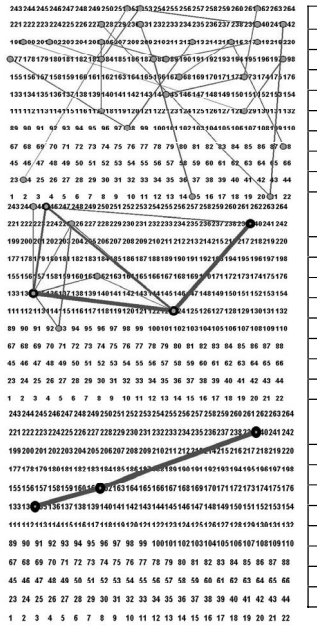
where $\tau_i(t)$ is the amount of pheromone trail for feature f_i (m/z window) at time t; η is a priori available information such as t-statistic or signal to noise ratio (SNR) for each feature; α and β are parameters which determine the relative influence of pheromone trail and a priori heuristic information, respectively.

At t=0, $\tau_i(t)$ is set to a constant for all features, allowing each feature to have equal probability of being selected. Thus, in the first iteration, ants choose randomly n distinct features (a trail) from L features. Let S_j be the jth ant consisting of n distinct features. Depending on the performance of S_j, the amount of pheromone trail for S_j will be updated. The performance function here is evaluated on the basis of disease state prediction capability of each S_j. We use the features in S_j to build an SVM classifier and estimate the prediction accuracy through the cross-validation method. The amount of pheromone trail for each feature in S_j is updated in proportion to prediction accuracy:

$$\tau_i(t+1) = \rho.\tau_i(t) + \Delta\tau_i(t)$$

where ρ is a constant between 0 and 1, representing the evaporation of pheromone trails. $\Delta\tau_i(t)$ is an amount proportional to the prediction accuracy achieved by S_j. $\Delta\tau_i(t)$ is set to zero, if $f_i \notin S_j$ at time t. This update is made for all M ants ($S_1,\ldots,S_M$). Note that at t=0, $\Delta\tau_i(t)$ is set zero for all features. The updating rule allows trails that yield good prediction to have their amount of pheromone trail increased, while others will evaporate. As the algorithm progresses, features with larger amounts of pheromone trails influence the probability function to lead the ants towards them.

To illustrate the ACO-SVM algorithm described above, we applied it to select three features from L=264. We used the SNR method proposed by Golub et al. (1999) as a priori heuristic information (η), α=β=1, and ρ=0.9. We define a feature as a location in the search space. Note that the dimension of the search space and the order of the features in the search space will not play a role, because the objective here is to maximize prediction accuracy, not distance between points. We placed the 264 features in a two-dimensional space where each location represents the labeled feature. M=10 ants were used to select n = three features. Initially, each ant chooses randomly three features (Fig. 2, top figure). The features selected are shown by the trails with three connected circles that lie on the selected features. At the 100th iteration, ants seem to favor some trails (middle figure). At the 284th iteration, all ants converged to one trail that goes through features 135, 162, and 240 (bottom figure). The prediction accuracy (found using the cross-validation method) improved from 79% at the 1st iteration to 91% at the 284th.



1st iteration			
Selected Sets			Accuracy %
118	189	206	79
231	188	15	78
240	24	252	75
.	.	.	.
.	.	.	.
262	173	145	55
168	88	216	50
100th iteration			
Selected Sets			Accuracy %
240	135	245	89
135	226	93	89
162	135	124	88
.	.	.	.
.	.	.	.
135	124	226	86
246	135	124	84
284th iteration			
Selected Sets			Accuracy %
240	162	135	91
240	162	135	91
240	162	135	91
.	.	.	.
.	.	.	.
240	162	135	91
240	162	135	91

Fig. 2. Pheromone trails for 10 ants at the 1st iteration (top), 100th iteration (middle) and 284th iteration (bottom).

III. Mass Spectral Data Analysis

A. Sample collection

Sample collection and generation of mass spectra was described previously [1, 2, 5]. The study examined an epidemic of viral infections in Egypt, a country where viral infections and associated HCC presents a serious health problem. The management of the disease would benefit from identification of biomarkers related to this disease. Serum samples of HCC cases and controls were obtained from Egypt between 2000 and 2002. Controls were recruited among patients from the orthopedic fracture clinic at the Kasr El-Aini Hospital (Cairo, Egypt) and were frequency-matched to cancer cases by gender, rural versus urban residency, and age [7]. Blood samples were collected in red top vacutainer tubes by trained phlebotomist each day around 10am and processed within a few hours according to a standard protocol. Aliquots of sera for mass spectrometric analysis were frozen at -80°C immediately after collection until analysis; all measurements were performed on samples of second-time thawed serum.

B. Sample preparation

The serum samples were enriched by denaturing ultrafiltration and desalting on C8 magnetic beads (MB) as described previously [1]. The procedure disrupts protein-protein interactions [8] and allows an efficient recovery of a Low Molecular Weight (LMW) serum fraction starting with 25 µl of serum. Eluted peptides were mixed with a matrix solution (3 mg/ml α-cyano-4-hydroxycinaminic acid in 50% actonitrile with 0.1% trifluoracetic acid), spotted onto

AnchorChip target (Bruker Daltonics, Billerica, MA) and analyzed using an Ultraflex MALDI TOF/TOF mass spectrometer (Bruker Daltonics). Each spectrum was detected in linear positive mode and was externally calibrated using a standard mixture of peptides. Denaturing ultrafiltration enriches LMW fraction of serum and plasma by removal of proteins greater than 50 kDa including albumin [1]. The enrichment improves quality of the spectra in the LMW region and allows analysis of approximately 300 peptides as described previously [1].

C. Data Preprocessing

Sixty-two replicate spectra were used to examine the run-to-run reproducibility of MALDI-TOF MS. Each spectrum consisted of about 136,000 m/z values with the corresponding ion intensities over the mass range 0.9 to 10 kDa. The dimension of the spectra was reduced to 23,846 m/z bins. A bin size of 100 ppm was found adequate. The mean of the intensities within each interval was used as the protein expression variable in each bin [9]. We transformed each intensity value by computing the base-two logarithm and found the mean log intensity value and standard deviation. The CV of the log-transformed intensity values in the 62 reference spectra ranged between 4.1% and 22.9% with a mean value of 10.5%.

For the remaining study, we used 84 HCC and 80 normal spectra. We excluded 14 spectra through outlier screening on the basis of their deviation from the median ion current, median record count (number of mass points), and their alignment with pre-selected landmarks. The remaining 150 spectra were binned, baseline corrected, and normalized. The baseline of each binned spectrum was estimated by obtaining the minimum value within a shifting window size of 50 bins. Spline approximation was used to regress the varying baseline. The regressed baseline was subtracted from the spectrum yielding a baseline corrected spectrum. We normalized each spectrum by dividing it its total ion current.

From the 150 preprocessed spectra (78 from patients with HCC and 72 from normal), we randomly selected 50 HCC and 50 normal (training spectra) for biomarker selection. The remaining 28 HCC and 22 normal (testing spectra) were set aside for later evaluation of the performance of the selected biomarkers.

We performed the following analyses using the 100 training spectra: (1) scaled the spectra to an overall maximum intensity of 100; (2) selected m/z values with reasonable intensity level and discarded those that appeared as noise, which was accomplished by identifying m/z values at which the slope sign changed from positive to negative and reasonable intensity was measured; (3) combined peaks if they differed in location by at most 7 bins or at most 0.03% relative mass. The method found 264 windows in the training spectra. For each spectrum, the maximum intensity within each window was found, yielding a 264 x 100 data matrix.

D. Biomarker Selection

We used the training spectra described in the previous section for biomarker (m/z window) selection. The validity of each classifier built with the selected biomarkers is evaluated using the sensitivity and specificity of the SVM classifier in distinguishing patients from healthy subjects. SVM classifiers are built for various combinations of m/z windows until the prediction accuracy of the SVM classifier converges or the maximum number of iteration is reached. The prediction accuracy is estimated through the four-fold cross-validation method.

To avoid any potential bias that may be introduced by parameter choice, the ACO-SVM algorithm was run for various numbers of features (n=3, 5, and 7) and ants (M=25, 50, and 100) with $\alpha=\beta=\gamma=1$, and $\rho=0.6$. Each combination (n features and M ants) was run 30 times, i.e., a total of 270 runs. Each run consisted of a maximum of 500 iterations. Figure 3 depicts the frequency of occurrence of the m/z windows in 270 runs. The figure suggests that the first seven m/z windows are frequently selected. Our TOF/TOF sequencing indicated that the first and the third m/z windows share the same sequence except for one amino acid. Thus, only the remaining six m/z windows are used in our subsequent analyses.

We used the SVM classifier to classify the testing spectra. We binned, baseline corrected, and normalized the testing spectra in the same way as the training spectra. Note that the testing spectra were scaled based on the parameters used to scale the training spectra. Figure 4 depicts the ROC curves and area under the ROC (AUROC) for the five markers both separately and combined by SVM. This figure demonstrates the advantage of a panel of biomarkers in achieving high prediction capability (100% sensitivity and 91% specificity) in distinguishing HCC patients from healthy individuals in the testing dataset.

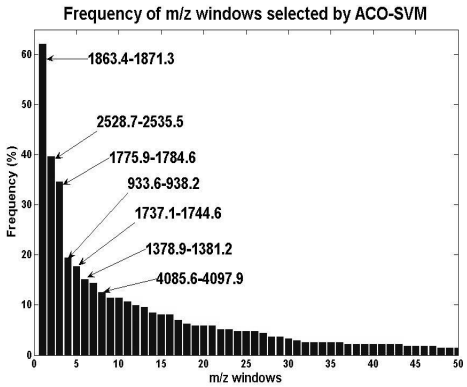


Fig. 3. Frequency of occurrence of m/z windows in 270 ACO-SVM runs sorted in decreasing order of frequency.

Finally, we compared three sets of features (the 23,846 m/z bins, 264 m/z windows, and the selected 6 m/z windows) in distinguishing HCC patients from healthy individuals using SVM classifiers. Note that each classifier was built using the training spectra and evaluated on the testing spectra. Figure

5 compares the ROC curves of the three SVM classifiers built using all bins, all m/z windows, and the six m/z windows. The figure shows that the AUROC for the SVM classifier with six m/z windows is larger than those that used all m/z bins or all m/z windows. Figure 6 shows the boxplots of the six m/z windows.

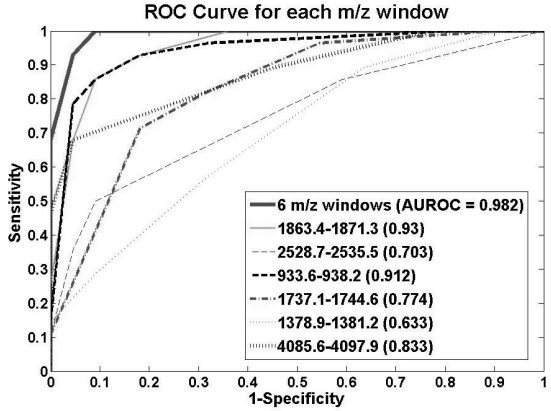


Fig. 4. ROC curves of each m/z window separately and all six combined. Note: the curves are based on testing spectra.

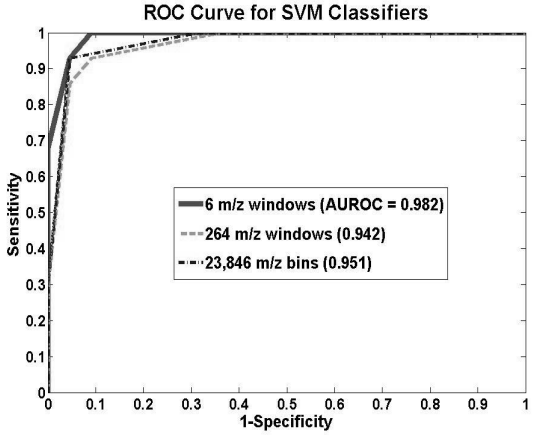


Fig. 5. ROC curves of three SVM classifiers (all bins, all m/z windows, and four m/z windows) on testing spectra.

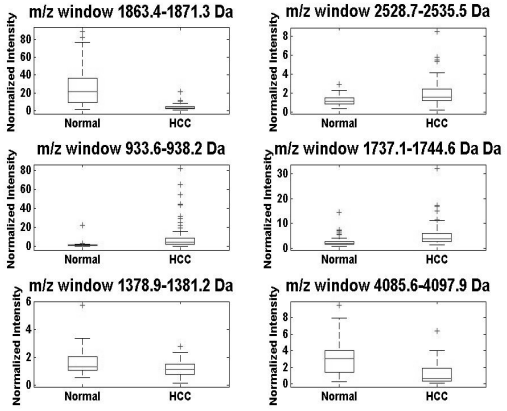


Fig. 6. Boxplots for the six m/z windows identified by the ACO-SVM (training and testing datasets combined).

IV. Conclusion

The high sensitivity and specificity achieved by the six candidate biomarkers indicate that MALDI-TOF technology, in conjunction with the proposed hybrid ACO-SVM algorithm could provide the means to discover novel biomarkers for HCC. The results also demonstrate the advantage of a panel of biomarkers in achieving high prediction capability.

Due to the initial trails which are determined randomly and the stochastic nature of the algorithm, every ACO-SVM run may not converge to the same trail in the search space. The frequency of occurrence of each m/z window in multiple runs allows us to estimate its relevance and the frequency response plot enables us to visually estimate the best number of m/z windows. Future work will focus on determining the frequency of occurrence of m/z windows that appear together (e.g. in pairs, triples, etc.) instead of combining the most frequent individual m/z windows via the frequency plot. The former will be useful to determine which m/z windows should be used together.

References

[1] E. Orvisky, S. K. Drake, B. M. Martin, H. Ressom, R. S. Varghese, D. Saha, G. L. Hortin, C. A. Loffredo, and R. Goldman, "Enrichment of low molecular weight fraction of serum for mass spectrometric analysis of peptides associated with hepatocellular carcinoma," *Proteomics In Press*, 2006.

[2] H. W. Ressom, R. S. Varghese, E. Orvisky, S. K. Drake, G. L. Hortin, M. Abdel-Hamid, C. A. Loffredo, and R. Goldman, "Analysis of MALDI-TOF serum profiles for biomarker selection and sample classification," *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2005.

[3] J. Villanueva, D. R. Shaffer, J. Philip, C. A. Chaparro, H. Erdjument-Bromage, A. B. Olshen, M. Fleisher, H. Lilja, E. Brogi, J. Boyd, M. Sanchez-Carbayo, E. C. Holland, C. Cordon-Cardo, H. I. Scher, and P. Tempst, "Differential exoprotease activities confer tumor-specific serum peptidome patterns," *J Clin Invest*, vol. 116, pp. 271-84, 2006.

[4] X. Zhang, S. M. Leung, C. R. Morris, and M. K. Shigenaga, "Evaluation of a novel, integrated approach using functionalized magnetic beads, bench-top MALDI-TOF-MS with prestructured sample supports, and pattern recognition software for profiling potential biomarkers in human plasma," *J Biomol Tech*, vol. 15, pp. 167-75, 2004.

[5] H. W. Ressom, R. S. Varghese, M. Abdel-Hamid, S. Abdel-Latif Eissa, D. Saha, L. Goldman, E. F. Petricoin, T. P. Conrads, T. D. Veenstra, C. A. Loffredo, and R. Goldman, "Analysis of mass spectral serum profiles for biomarker selection," *Bioinformatics* vol. 21, pp. 4039-4045, 2005.

[6] M. Dorigo, G. Di Caro, and L. M. Gambardella, "Ant algorithms for discrete optimization," *Artif Life*, vol. 5, pp. 137-72, 1999.

[7] S. Ezzat, M. Abdel-Hamid, S. Abdel-Latif Eissa, N. Mokhtar, N. A. Labib, L. El-Ghorory, N. N. Mikhail, A. Abdel-Hamid, T. Hifnawy, G. T. Strickland, and C. A. Loffredo, "Associations of pesticides, HCV, HBV, and hepatocellular carcinoma in Egypt," *Int J Hygiene Env Health, in press*, 2005.

[8] R. S. Tirumalai, K. C. Chan, D. A. Prieto, H. J. Issaq, T. P. Conrads, and T. D. Veenstra, "Characterization of the low molecular weight human serum proteome," *Mol Cell Proteomics*, vol. 2, pp. 1096-103, 2003.

[9] J. Villanueva, J. Philip, D. Entenberg, C. A. Chaparro, M. K. Tanwar, E. C. Holland, and P. Tempst, "Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry," *Anal Chem*, vol. 76, pp. 1560-70, 2004.