Rapid Mass Spectrometric Metabolic Profiling of Blood Sera Detects Ovarian Cancer with High Accuracy

Manshui Zhou1, Wei Guan3, L. DeEtte Walker2,4, Roman Mezencev2,4, Benedict B. Benigno4, Alexander Gray3, Facundo M. Fernández1, and John F. McDonald2,4

Abstract

Background: Ovarian cancer diagnosis is problematic because the disease is typically asymptomatic, especially at the early stages of progression and/or recurrence. We report here the integration of a new mass spectrometric technology with a novel support vector machine computational method for use in cancer diagnostics, and describe the application of the method to ovarian cancer.

Methods: We coupled a high-throughput ambient ionization technique for mass spectrometry (direct analysis in real-time mass spectrometry) to profile relative metabolite levels in sera from 44 women diagnosed with serous papillary ovarian cancer (stages I-IV) and 50 healthy women or women with benign conditions. The profiles were input to a customized functional support vector machine–based machine-learning algorithm for diagnostic classification. Performance was evaluated through a 64-30 split validation test and with a stringent series of leave-one-out cross-validations.

Results: The assay distinguished between the cancer and control groups with an unprecedented 99% to 100% accuracy (100% sensitivity and 100% specificity by the 64-30 split validation test; 100% sensitivity and 98% specificity by leave-one-out cross-validations).

Conclusion: The method has significant clinical potential as a cancer diagnostic tool. Because of the extremely low prevalence of ovarian cancer in the general population (~0.04%), extensive prospective testing will be required to evaluate the test’s potential utility in general screening applications. However, more immediate applications might be as a diagnostic tool in higher-risk groups or to monitor cancer recurrence after therapeutic treatment.

Impact: The ability to accurately and inexpensively diagnose ovarian cancer will have a significant positive effect on ovarian cancer treatment and outcome. Cancer Epidemiol Biomarkers Prev; 19(9); OF1–10. ©2010 AACR.

Introduction

Ovarian cancer (OC) is the most lethal of the gynecologic cancers and is the fifth leading cause of all cancer-related deaths among women (1). Although the 5-year survival rate for women diagnosed with the disease early in its progression is >90%, the survival rate for patients diagnosed at later stages is only ~20% (2). The main challenge with OC is that it typically arises and progresses initially without well-defined clinical symptoms (3). Thus, successful diagnosis plays a central role in deciding appropriate therapy and improving patient prognosis.

Most OC blood tests in current clinical practice monitor level changes of a single molecule that has been shown to be elevated (or lowered) in a significant number of diseased patients. Although these tests are often not definitive per se, they may be of significant predictive value when combined with other procedures. However, single molecule-based OC diagnostic assays have had only limited diagnostic power. More recent research has focused on tests based on panels of biomarkers. For example, a recently developed test that looks at six serum proteins has been shown to be of significant diagnostic value in high ovarian cancer risk groups (e.g., BRCA1-positive patients; refs. 4, 5).

We report here on the coupling of a high-throughput ambient ionization technique for mass spectrometry (MS) with machine-learning approaches for the metabolomic classification of sera from ovarian cancer and control patients. This technique, known as direct analysis in real-time (DART; ref. 6), is one of the members of the rapidly growing family of open-air (ambient) ionization...
methods for MS (7) that also includes desorption electrospray ionization (8). We view this as a successful step towards the development of an accurate new approach for the diagnosis of ovarian and other cancers. In this DART MS test, a stream of excited metastables is used to desorb and chemically ionize a dried drop of derivatized serum (Fig. 1). A typical DART MS profile displays a multitude of signals corresponding to metabolites rapidly desorbed and ionized in a time-dependent fashion (Fig. 1, c.x.). These profiles are then used as input for a customized functional support vector machine (fSVM)-based machine-learning algorithm for the classification of serum samples. The assay distinguished between the cancer and control groups with 99% to 100% accuracy (100% sensitivity and 98-100% specificity) under two different data evaluation approaches.

Figure 1. Diagram of study design and workflow showing metabolomic investigation of serum samples for detection of ovarian cancer by DART-TOF MS.

A, serum sample preparation: (i) protein precipitation, centrifugation, and separation of the metabolite-containing supernatant followed by (ii) evaporation of solvent to generate a metabolite-containing pellet. This pellet is then subjected to (iii) derivatization to increase the volatility of polar metabolites.

B, schematic of the DART-TOF MS equipped with a custom-built sample arm: (iv) glow discharge compartment, (v) gas heater, and (vi) ionization region where the sample-carrying capillary is placed. Differentially pumped atmospheric pressure interface (vii) to transport ions towards the mass analyzer. Radiofrequency ion guide (viii) where ions are collisionally cooled prior to entering the orthogonal TOF mass analyzer (ix). C, typical data are acquired in a time-resolved fashion [x, three-dimensional contour plots of single runs corresponding to an ovarian cancer patient (top) and a control (bottom)]. The region of the time-resolved signal with best signal-to-noise ratio was averaged yielding profile mass spectra (xi) reflecting metabolic fingerprints. D, machine-learning techniques such as SVMs are used for building a multivariate classifier (xii, objects in original variable space; xiii, objects in classifier space).
Sample collection

Serum samples were obtained from the Ovarian Cancer Institute laboratory at Georgia Tech after approval by the Institutional Review Board from Northside Hospital and Georgia Institute of Technology (Atlanta, GA; Table 1). All donors were required to fast and to avoid medicine and alcohol for 12 hours prior to sampling (except for certain allowable medications, for instance, diabetics were allowed insulin). Following informed consent by donors, 5 mL of whole blood were collected by venipuncture into evacuated blood collection tubes that contained no anticoagulant (blood taken prior to the administration of anesthesia, immediately preceding surgery). Within 1 hour of venipuncture, serum was collected and 200-μL aliquots of each sample were stored in 1.5 mL microtubes at −80°C until ready to use.

Sample preparation

Prior to analysis, 200 μL of each serum sample were thawed on ice and mixed with 1 mL of freshly prepared, chilled (−18°C), and degassed 2:1 (v/v) acetone/isopropanol mixture. The mixture was vortexed and placed in a freezer at −18°C overnight to precipitate proteins followed by centrifugation at 13,000 × g for 5 minutes. The supernatant was transferred to a new centrifuge tube, and the solvent was evaporated in a speed vacuum. The solid residue was redissolved in 25 μL of anhydrous pyridine (EMD Chemicals), and shaken for 1 hour at room temperature for complete dissolution. Fifty microliters of N-trimethylsilyl-N-methyltrifluoroacetamide (Alfa Aesar) containing 0.1% trimethylchlorosilane (Alfa Aesar) were added to the sample in a N2-purged glove box. The mixture was then incubated at 50°C in an inert N2 atmosphere for half an hour, resulting in trimethylsilane derivatization of amide, amine, carboxyl, and hydroxyl groups. The final derivatized mixture was subject to DART MS analysis.

DART-TOF MS

A in-depth characterization of the analytical figures of merit of the DART MS approach used here has been recently reported (9), and therefore, the method is only briefly presented. Serum mass spectrometric analysis was done using a DART ion source (IonSense, Inc.) coupled to a JEOL AccuTOF orthogonal time-of-flight (TOF) MS (JEOL, Inc., Japan). Prior to DART MS analysis, 0.5 μL of derivatized serum solution was pipette-deposited onto the glass end of the Dip-tip applicator (IonSense) coupled to the sampling arm, a 1.2-minute data acquisition run was started, and the sample allowed to air-dry for 0.65 minutes. The sampling arm was then rapidly switched so that the dried sample was exposed to the ionizing zone of the DART ion source. After 0.9 minutes in the acquisition run (0.25-minute sampling time), the sample was removed, and a new Dip-tip placed on the sample holder while the remaining 0.3 minutes of the run was completed. Each sample was run in triplicate.

The DART ion source was operated in positive ion mode with a helium gas flow rate of 3.0 L/min heated to 200°C. The glass tip-end was positioned 1.5 mm below the MS inlet. The discharge needle voltage of the DART source was set to +3,600 V, and the perforated, and grid electrode voltages set to +150 and +250 V, respectively. Accurate mass spectra were acquired in the range of m/z 60 to 1,000 with a spectral recording interval of 1.0 seconds, and an RF ion guide peak voltage of 1,200 V. The settings for the TOF MS were as follows: ring lens, +8 V; orifice 1, +40 V; orifice 2, +6 V; orifice 1 temperature, 80°C; and detector voltage, −2,800 V. Mass drift compensation was done after analysis of each sample using a 0.20 mmol/L polyethylene glycol 600 standard (PEG 600, Fluka Chemical Corp.) in methanol. The measured resolving power of the TOF MS was 6,000 at full-width, half-maximum, with observed mass accuracies in the range of 2 to 20 ppm, depending on the signal-to-noise ratios of the particular peak investigated. A repeatability of 4.1% to 4.5% was obtained for the total ion signal using a manual sampling arm.

Data preprocessing

All profile mass spectra were obtained by time-averaging of the total ion chronogram between 0.73 and 0.76 minutes after each injection, which corresponds to the part of the time-varying signal that is conducive to the maximum number of analytes detected and identified with good sensitivity (9). Following DART-TOF MS data collection and mass drift compensation, the background spectrum was subtracted and profile spectral data were exported in JEOL-DX format and converted to a comma-separated format prior to importing in MATLAB 7.6.0 (R2008a, MathWorks). The resulting data were normalized to a relative intensity scale and re-sampled to a total of 20,000 features between m/z 60 and 990 using the msresample function in the Matlab Bioinformatics Toolbox (10). The three replicate DART spectra were then averaged. The original data set containing the DART-MS data can be downloaded (11).

Data analysis

Evaluation framework. The prediction performance of the data set was evaluated through a 64-30 split validation without feature selection, and with different feature selection methods. Following this approach, further evaluation of the classifier performance was done through leave-one-out cross-validation (LOOCV). In each case, the chosen feature selection method was applied only to the training data set, and then the prediction performance of the selected feature subset on the test data set was measured.

SVM (12) analysis of averaged DART mass spectra was done using libSVM (13). fSVM analysis was done using the functional data analysis package (14) and libSVM. Partial least squares discriminant analysis was done using the PLS Toolbox (version 4.1, Eigenvector Research). We implemented ANOVA, recursive feature elimination
Table 1. Patients analyzed in this study

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Ovarian histopathology</th>
<th>Stage/grade</th>
<th>Age at surgery</th>
<th>Menopause status</th>
<th>Endometriotic cysts present?</th>
</tr>
</thead>
<tbody>
<tr>
<td>242</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/3</td>
<td>63</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>281</td>
<td>Papillary serous carcinoma</td>
<td>III/1</td>
<td>66</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>454</td>
<td>Papillary serous carcinoma</td>
<td>III/3</td>
<td>72</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>458</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/3</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>472</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/2-3</td>
<td>49</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>473</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/3</td>
<td>48</td>
<td>Perimenopausal</td>
<td>No</td>
</tr>
<tr>
<td>491</td>
<td>Papillary serous carcinoma</td>
<td>I/</td>
<td>74</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>495</td>
<td>Papillary serous carcinoma</td>
<td>IIIb/3</td>
<td>43</td>
<td>Premenopausal</td>
<td>No</td>
</tr>
<tr>
<td>512</td>
<td>Papillary serous carcinoma</td>
<td>IIIb/2-3</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>517</td>
<td>Papillary serous carcinoma</td>
<td>Ia/3</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>526</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/2-3</td>
<td>49</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>528</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/3</td>
<td>66</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>529</td>
<td>Papillary serous carcinoma</td>
<td>IIIc</td>
<td>67</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>533</td>
<td>Papillary serous carcinoma</td>
<td>III/1</td>
<td>43</td>
<td>Premenopausal</td>
<td>No</td>
</tr>
<tr>
<td>537</td>
<td>Papillary serous carcinoma</td>
<td>IIIa/2-3</td>
<td>64</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>542</td>
<td>Papillary serous carcinoma</td>
<td>IV/3</td>
<td>61</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>551</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/IV/3</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>555</td>
<td>Papillary serous carcinoma</td>
<td>IV/3</td>
<td>49</td>
<td>Perimenopausal</td>
<td>No</td>
</tr>
<tr>
<td>588</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/2-3</td>
<td>71</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>589</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/IV/3</td>
<td>46</td>
<td>Premenopausal</td>
<td>No</td>
</tr>
<tr>
<td>606</td>
<td>Papillary serous carcinoma</td>
<td>IIIa/3</td>
<td>54</td>
<td>Premenopausal</td>
<td>No</td>
</tr>
<tr>
<td>617</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/2-3</td>
<td>64</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>620</td>
<td>Papillary serous carcinoma</td>
<td>III/IV/3</td>
<td>62</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>632</td>
<td>Papillary serous carcinoma</td>
<td>IIIb/3</td>
<td>65</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>643</td>
<td>Papillary serous carcinoma</td>
<td>IIIb/2</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>644</td>
<td>Papillary serous carcinoma</td>
<td>IIIb/1-2</td>
<td>46</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>647</td>
<td>Papillary serous carcinoma</td>
<td>IIIb-c/3</td>
<td>68</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>651</td>
<td>Papillary serous carcinoma</td>
<td>IIIb-c/3</td>
<td>46</td>
<td>Perimenopausal</td>
<td>No</td>
</tr>
<tr>
<td>655</td>
<td>Papillary serous carcinoma</td>
<td>III/IV/3</td>
<td>75</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>659</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/IV/3</td>
<td>78</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>678</td>
<td>Papillary serous carcinoma</td>
<td>IV/3</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>688</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/3</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>694</td>
<td>Papillary serous carcinoma</td>
<td>IIIb/3</td>
<td>70</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>704</td>
<td>Papillary serous carcinoma</td>
<td>III/IV/3</td>
<td>75</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>717</td>
<td>Papillary serous carcinoma</td>
<td>IIIb/3</td>
<td>64</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>721</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/1</td>
<td>58</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>756</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/2</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>782</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/3</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>787</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/3</td>
<td>72</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>821</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/1-2</td>
<td>58</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>831</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/</td>
<td>69</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>864</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/3</td>
<td>60</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>876</td>
<td>Papillary serous carcinoma</td>
<td>IIIa/1</td>
<td>63</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>5010</td>
<td>Papillary serous carcinoma</td>
<td>IIIc/1</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>440</td>
<td>Within normal limits</td>
<td>N/A</td>
<td>50</td>
<td>Perimenopausal</td>
<td>No</td>
</tr>
<tr>
<td>504</td>
<td>Within normal limits</td>
<td>N/A</td>
<td>48</td>
<td>Premenopausal</td>
<td>No</td>
</tr>
<tr>
<td>523</td>
<td>Serous cystadenoma</td>
<td>N/A</td>
<td>32</td>
<td>Premenopausal</td>
<td>No</td>
</tr>
<tr>
<td>534</td>
<td>Within normal limits</td>
<td>N/A</td>
<td>72</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>540</td>
<td>Within normal limits</td>
<td>N/A</td>
<td>59</td>
<td>Postmenopausal</td>
<td>No</td>
</tr>
<tr>
<td>541</td>
<td>Within normal limits</td>
<td>N/A</td>
<td>41</td>
<td>Perimenopausal</td>
<td>No</td>
</tr>
</tbody>
</table>

(Continued on the following page)
(15), and Weston’s (16) feature selection methods in Matlab 7.6.0. Mangasarian’s L1-norm SVM (17, 18) was also implemented in Matlab.

\textbf{fSVM}. In some application domains such as chemometrics, it is well known that the shape of a spectrum is sometimes more important than its actual mean value. Therefore, it is beneficial to view the intensity as functions of \( m/z \) values, and perform functional classification. The goal of functional classification (19) is to predict the label \( y \) of a functional data instance \( X \) given training data

\[
S = \{X_i, y_i\}_{i=1}^{M}, \quad X_i \in H,
\]

where the input functional data instance \( X_i \) is a random variable that takes values

\begin{table}
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
Patient ID & Ovarian histopathology & Stage/grade & Age at surgery & Menopause status & Endometriotic cysts present? \\
\hline
544 & Within normal limits & N/A & 49 & Perimenopausal & No \\
552 & Within normal limits & N/A & 41 & Premenopausal & No \\
612 & Within normal limits & N/A & 48 & Premenopausal & No \\
614 & Within normal limits & N/A & 44 & Premenopausal & No \\
615 & Within normal limits & N/A & 42 & Perimenopausal & No \\
623 & Simple cyst & N/A & 54 & Postmenopausal & No \\
627 & Within normal limits & N/A & 59 & Postmenopausal & No \\
636 & Within normal limits & N/A & 71 & Postmenopausal & Yes \\
650 & Cystic corpus luteum & N/A & 47 & Postmenopausal & No \\
677 & Within normal limits & N/A & 68 & Postmenopausal & No \\
691 & Within normal limits & N/A & 70 & Postmenopausal & No \\
693 & Simple cyst & N/A & 60 & Postmenopausal & No \\
697 & Within normal limits & N/A & 51 & Premenopausal & Yes \\
698 & Functional cyst & N/A & 49 & Perimenopausal & No \\
703 & Simple cyst & N/A & 42 & Premenopausal & No \\
719 & Within normal limits & N/A & 55 & Premenopausal & No \\
733 & Within normal limits & N/A & 37 & Postmenopausal & No \\
736 & Within normal limits & N/A & 45 & Premenopausal & No \\
737 & Within normal limits & N/A & 41 & Perimenopausal & No \\
740 & Functional cyst & N/A & 37 & Premenopausal & No \\
749 & Simple cyst/cystic follicles & N/A & 56 & Perimenopausal & No \\
750 & Serous cystadenoma & N/A & 41 & Postmenopausal & No \\
751 & Within normal limits & N/A & 60 & Postmenopausal & No \\
752 & Within normal limits & N/A & 74 & Postmenopausal & No \\
755 & Within normal limits & N/A & 75 & Postmenopausal & No \\
757 & Within normal limits & N/A & 84 & Postmenopausal & No \\
759 & Within normal limits & N/A & 52 & Postmenopausal & No \\
763 & Hemorrhagic cyst & N/A & 45 & Premenopausal & No \\
765 & Within normal limits & N/A & 84 & Postmenopausal & No \\
766 & Within normal limits & N/A & 36 & Premenopausal & No \\
783 & Within normal limits & N/A & 52 & Premenopausal & No \\
790 & Cystic follicles & N/A & 39 & Premenopausal & No \\
796 & Within normal limits & N/A & 44 & Premenopausal & No \\
808 & Within normal limits & N/A & 35 & Premenopausal & No \\
828 & Simple cyst & N/A & 59 & Postmenopausal & No \\
829 & Simple cyst & N/A & 33 & Postmenopausal & No \\
838 & Within normal limits & N/A & 51 & Perimenopausal & No \\
839 & Simple cyst & N/A & 79 & Postmenopausal & Yes \\
842 & Within normal limits & N/A & 70 & Postmenopausal & Yes \\
846 & Hemorrhagic corpus luteum & N/A & 51 & Premenopausal & No \\
848 & Within normal limits & N/A & 70 & Postmenopausal & No \\
NHS1 & Healthy serum donor & N/A & 36 & Premenopausal & No \\
NHS4 & Healthy serum donor & N/A & 34 & Premenopausal & No \\
NHS10 & Healthy serum donor & N/A & 37 & Premenopausal & No \\
\hline
\end{tabular}
\caption{Patients analyzed in this study (Cont’d)}
\end{table}
in an infinite dimensional Hilbert space $H$, the space of functions.

In practice, the functions that describe the input data instance $X_1$, ..., $X_M$ are never perfectly known. Often, $N$ discretization points have been chosen in $t_1$, ..., $t_N \in R$, and each functional data instance $X_i$ is described by a vector $(X_i(t_1), ..., X_i(t_N)) \in R^N$. Sometimes, the functional data instances are badly sampled and the number and the location of discretization points are different between different instances. The usual solution under this context is to construct an approximation (such as B-spline interpolation) for each instance of $X_i$ based on its observation values, and then apply sampling uniformly to the reconstructed functional data (20). Therefore, a simple solution is to apply the standard SVM to the vector representation of the functional data. With the introduction of functional transformations and functional kernels, it has been suggested recently that the classification accuracy could be improved by designing SVMs specifically for functional classification, with the introduction of functional transformations and function kernels (21, 22):

1. Apply functional transformation, projection $P_{\Psi}$, on each instance $X_i$ as $P_{\Psi}(X_i) = x_i = (x_{i1}, ..., x_{iN})$ with $X_i$ approximated by $\sum_{k=1}^{\infty} x_{ik} \Psi_k$, where $\\{\Psi_k\}_{k \geq 1}$ is a complete orthonormal basis of the functional space $H$.
2. Build a standard SVM on the coefficients $x_i \in R^N$ for all $i = 1, ..., M$.

This procedure is equivalent to working with a functional kernel, $K_{\Psi}(x_i, x_j)$ defined as $K(P_{\Psi}(X_i), P_{\Psi}(X_j))$, where $P_{\Psi}$ denotes the projection onto the $N$-dimensional subspace $V^N \subset H$ spanned by $\\{\Psi_k\}_{k=1}^{\infty}$, and $K$ denotes any standard SVM kernel.

Good candidates for the basis functions include the Fourier basis and wavelet basis. If the functional data are known to be nonstationary, a wavelet basis might yield better results than the Fourier basis (20). Other suitable choices include B-spline bases, that generally perform well in practice (21).

Feature selection. The ANOVA is one of the most commonly used filter-based feature selection methods in bioinformatics. It helps to identify the features that highlight differences between groups (23, 24). Let the data set $S$ contain $c$ classes (groups), $n$ data instances, and $n_i$ instances from each class $c_i$, $X_{ij}$ ($i = 1, ..., c; j = 1, ..., n_i$) be a random sample of size $n_i$ from a population with mean $u_i$. ANOVA is used to investigate the null hypothesis $H_0$: $u_1 = u_2 = ... = u_c$ through F-test $F = \frac{SSB/(c-1)}{SSW/(n-c)}$ where $SSB = \sum_{i=1}^{c} (\bar{x}_i - \bar{x})^2$ is the interclass sum of square, $SSW = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ is the total intra-class sum of squares, $\bar{x}_i$ and $\bar{x}$ are estimates of class and overall sample means, respectively; $x_{ij}$ is an observation (sample) from class $c_i$. If the null hypothesis is rejected [$F > F_{c-1,n-c,\alpha}$] (the upper 100th percentile of the $F$ distribution with $c-1$ and $n-c$ degrees of freedom), this implies that the groups of data samples differ significantly.

Elemental formula determination and metabolic database matching

Features in the SVM model using 1:7:20,000 subsampling were assigned elemental formulae and tentatively matched to metabolites by finding the closest mass spectral peak matching the model features in the 103 to 714 $m/z$ range. This $m/z$ range was chosen because it is fully covered by the TOF calibration function thus providing the most reliable accurate masses. No attempt was made to match SVM model features outside this range. Accurate masses were searched against a custom-built database containing 2,924 entries corresponding to elemental formulae of endogenous human metabolites in the Human Metabolome Database (25). Each entry was manually expanded to take into account the mono-trimethylsilyl, di-trimethylsilyl, and/or tri-trimethylsilyl derivatives. Entries for families of compounds not reacting with the N-trimethylsilyl-N-methyltrifluoroacetamide/triethylchlorosilane reagent mixture were not expanded. Matching of database records to experimental DART MS data was done using the SearchFromList application part of the Mass Spec Tools suite of programs (ChemSW) using a tolerance of 10 mnu to obtain candidate elemental formulae. If no matches were found, the next closest match within 20 mnu was selected.

Results

Metabolic profiles can distinguish between ovarian cancer and control samples

Serum samples were obtained from 44 women diagnosed with serous papillary ovarian cancer (stages I-IV) and 50 healthy women or women with benign conditions (e.g., serous, simple, or follicular cysts; Table 1) and subjected in triplicate to DART MS profiling.

We have previously tested a customized SVM algorithm for the classification of metabolic profiles obtained by using liquid chromatography-MS (26). In this study, the classification procedure builds on our previous work and can be briefly described as follows: (a) the data are collapsed along the desorption time dimension by using the average value within the time range of interest for all mass spectral $m/z$ values (“features”); (b) the resulting feature vector is smoothed using B-splines (12, 27) to create the functional representation; (c) the vector of spline coefficients is then used by the SVM (17). To deal with the very large number of features (20,000 $m/z$ values per serum sample run), the data were subjected to a variety of data analysis methods including SVM (the above described nonlinear ISVM as well as the standard linear SVM; ref. 28), and partial least squares discrimination analysis (29, 30). Classification was done either with all mass spectral features, or with feature subsets selected by simple subsampling (15, 16).

We evaluated the efficacy of our classifiers by two separate approaches. In both approaches, a “training set” of samples was used to build the predictive model and a separate and independent set of samples (“test set”) was
used to evaluate the predictive/discriminatory power of the model, and the best model parameters to avoid overfitting or underfitting. A summary of the results obtained with fSVMs is presented in Table 2. In the first approach, we used a training set of 64 patients selected at random, with the 30 remaining patients treated as an independent test set (64-30 split validation). In this case, the fSVM with linear kernels applied to a feature subset selected by one-way ANOVA \((P \leq 0.05)\) achieved 100% accuracy (100% sensitivity and 100% specificity; Table 2A; Fig. 2A).

Although the above approach is considered the “gold standard” in evaluating diagnostic tests (4), LOOCV is a more rigorous approach to model parameter estimation due to its maximal usage of the data for training (31, 32). During LOOCV, each training set consisted of all patient samples except for one “left out” sample that is tested. In this way, each one of the patient samples is sequentially treated as an unknown, classified by the model as “cancer” or “control” in a blind fashion, and the accuracy of each classification evaluated. While validating models by LOOCV, feature selection was done independently on 94 different 94-1 \((n-1)\) split validations. Only one 94-1 split validation resulted in a misclassification giving an overall accuracy of 98.9% (100% sensitivity and 98% specificity; Table 2B; Fig. 2B). For comparison purposes, standard SVMs and partial least squares discrimination analyses were also used to classify the data. The corresponding results are presented in Supplementary Table S1. All classification and feature selection methods showed high accuracy owing to the inherent discriminative power of the data, but the best performance was obtained using fSVM.

**Pathway enrichment analysis and metabolic network building**

The MetaCore 5.2 (GeneGO) software suite was used for metabolic network analysis. One hundred and fifty-three estimated elemental formulae (Supplementary Table S2) obtained by DART MS accurate mass measurements of differentiating spectral features were assigned to 385 network objects by the metabolic network analysis software, of which 299 represented unique endogenous metabolites or xenobiotic compounds (33). Metabolic compounds assigned to these elemental formulae by MetaCore were mapped onto GeneGO canonical metabolic pathways that were ranked according to their relevance to the input set using \(P\) values calculated based on hypergeometric distribution. These differentiating compounds mapped onto 25 pathways (34) with \(P < 0.01\) (Supplementary Fig. S1), suggesting differences between cancer and noncancer groups in amine, amino acid, eicosanoid, and TTP metabolisms. Suggested differences in the metabolisms of carbohydrates and androgens/estrogens have lower confidence because the relevance of corresponding pathways was determined from ambiguously identified metabolites (e.g., several different hexoses corresponding with elemental formula \(C_{n}H_{2}O_{6}\)) and were not further examined.

**Discussion**

**Potential biological significance of metabolic changes in ovarian cancer**

A considerable proportion of the differentiating metabolites identified during the development of our assay represents components of the histamine pathway (Supplementary Figs. S2 and S5). Serum histamine levels have also been reported to be altered in breast cancer (35). Histamine is known to serve as a receptor-dependent growth factor in some colon, gastric, breast cancer, and melanoma cell lines and to inhibit lymphocyte responsiveness via proliferation and activation of T lymphocyte suppressor cells (36). In addition, the relationship of histamine with the metabolism of nitric oxide, polyamines, and angiogenesis is an emerging area of interest in cancer biology (37). The overrepresentation of members of the histamine pathway in our metabolic panel suggests that these species might also be of functional importance in ovarian cancer.

Other pathways overrepresented in our data set suggest that changes in the metabolism of several amino acids (e.g., glycine) involved in the de novo synthesis of purine nucleotides are also altered in ovarian cancer. Glycine, serine, and sarcosine were all tentatively identified as differentiating metabolites in our study and these metabolites

---

**Table 2. Results from the analysis of DART MS ovarian cancer data using fSVMs**

<table>
<thead>
<tr>
<th>(A) Classifier type</th>
<th>Feature selection method</th>
<th>No. of features</th>
<th>SENS (%)</th>
<th>SPEC (%)</th>
<th>ACC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fSVM</td>
<td>One-way ANOVA</td>
<td>3,017</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

| (B) fSVM            | One-way ANOVA            | 4,390          | 100      | 98       | 98.9    |

NOTE: ANOVA feature selection in combination with fSVM was first applied to the training data set and then the test set predicted using the selected features subset: (A) 64-30 split validation, (B) LOOCV evaluation. The sensitivity (SENS), specificity (SPEC), and accuracy (ACC) were determined as follows: SENS, true positive (TP)/TP + false negatives (FN); SPEC, true negative (TN)/TN + false positives (FP); ACC, (TP + TN)/(TP + FN + TN + FP).
are components of overrepresented canonical pathways of alanine, serine, cysteine, threonine, and glycine metabolisms (Supplementary Figs. S3 and S5). Several amino acids from these pathways have previously been identified in an earlier MS-based metabolic profile of ovarian cancer tissues (38). Sarcosine, the N-methyl derivative of glycin, is elevated in invasive prostate cancer cell lines and in the tumors and urine of patients with metastatic...
prostate cancer (39). Also consistent with our findings, the levels of these amino acids have all been previously reported to be elevated in the sera of patients with colorectal (40), lung, and breast (41) cancers.

A number of other tentatively identified metabolites (e.g., dopamine, tyramine, 5-hydroxykynurenamine, and 1,2-dehydrosalsolinol) which are differentially expressed in the sera of ovarian cancer relative to control patients are all products of decarboxylation of their precursor amino acids catalyzed by aromatic L-amino acid decarboxylase (DDC). This enzyme and its metabolic products have previously been shown to be elevated in neuroendocrine neoplastic tissues (carcinoid, small cell lung cancer; ref. 41). We have recently reported that DDC is overexpressed in ovarian cancer (42). Networks built from our metabolic data set using dopamine, tyramine, 5-hydroxykynurenamine, and 1,2-dehydrosalsolinol and their precursors (Supplementary Figs. S4 and S5) are consistent with the finding that DDC (and its metabolic products) is (are) differentially expressed in ovarian cancer.

The utility of metabolic profiling as a diagnostic test for ovarian cancer

Previous efforts to discover more accurate biomarkers of ovarian cancer using MS have generally focused on large biopolymers, such as proteins (43). However, finding and validating biomarkers of this kind has been plagued by the fact that the serum proteome is extremely complex, comprising ∼2 × 10^5 protein species with a dynamic range spanning 10 orders of magnitude (44). This inherent complexity, combined with current limitations in the proteomic analytical toolbox, could result in the convolution of biomarker variability with nonbiological sources of variance. Comprised of ∼2,500 molecules with molecular weights of <1,000 Da, the known components of the serum metabolome could readily be distinguished from the serum proteome and more thoroughly investigated (45). As biological studies using more sensitive analytical tools with higher peak capacities improve our understanding of the serum metabolome, the number of detected and identified metabolites is expected to progressively increase, enriching the biological significance of discriminating spectral features useful in diagnostics.

MS analysis of serum samples typically employs chromatographic separation. This step is usually time-consuming and could result in increased costs and memory effects, which we believe was one of the confounding factors in our previous liquid chromatography-MS study (26). Our DART method circumvents chromatographic separation, making use of direct ionization without a matrix in a noncontact fashion. This decreases cross-contamination between experiments, enabling a better detection of differences between disease and control groups. Moreover, DART is able to ionize a broad range of metabolites with varying polarities (46), allowing simultaneous interrogation of multiple chemical species at minimal cost.

By combining the DART-TOF MS with a customized fSVM classification algorithm, we were able to distinguish sera from cancer patients and controls with 100% accuracy as estimated by the 64-30 split validation test, as well as 99% accuracy using the more stringent LOOCV test (100% sensitivity and 98% specificity). In this study, the use of high-resolution TOF MS was necessary for metabolite identification purposes, but the spectral data were later down-sampled for machine-learning purposes, suggesting that approaches similar to the one presented here, but based on low-resolution MS data acquisition, might also be conducive to high discriminatory power.

There is a general consensus among the ovarian cancer community that to be of clinical significance, a diagnostic test for ovarian cancer must have a minimum positive predictive value of ~10% (47). Because the prevalence of ovarian cancer in the general population is low (~0.04%), the accuracy of any potential screening test to be used in the general population must be extremely high (~100%, ref. 3). Although our results indicate that our approach has great potential as a diagnostic tool of clinical significance, more extensive testing will be required to define its use in screening applications. Other, more immediate clinical applications of our assay may be in those subpopulations of women in which the prevalence of ovarian cancer is known to be relatively high. For example, the estimated incidence of ovarian cancer in women ages 20 and over with two first-degree relatives with ovarian cancer is 0.266% (48). Using incidence to approximate prevalence (49), we estimate a clinically significant 12% positive predictive value for our assay in this subpopulation, assuming the more stringent LOOCV values of 100% sensitivity and 98% specificity. Women 20 years of age and over who test positive for BRCA1 or BRCA2 are reported to have an incidence of ovarian cancer as high as 0.683% (50). For this group of women, our assay would have an estimated positive predictive value of 26%—well above the minimum value (~10%) for a test to be considered of clinical significance.

The results presented here show the potential application of our method as an ovarian cancer diagnostic of significant clinical value. In addition, if future studies establish that metabolic profiles of different cancers and other diseases are sufficiently distinct, our method might have the added advantage that it could be used to rapidly and inexpensively test for multiple diseases from a small serum sample.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Grant Support

Georgia Research Alliance/VentureLabs program (F.M. Fernández and J.F. McDonald), a Blanchard Professorship (F.M. Fernández), the Deborah Nash Harris Endowment Fund (J.F. McDonald), Northside Hospital (Atlanta; J.F. McDonald and B.B. Benigno), The Ovarian Cycle Fund (J.F. McDonald and B.B. Benigno), and the Robinson Family Foundation Fund (J.F. McDonald).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 02/02/2010; revised 06/03/2010; accepted 06/24/2010; published OnlineFirst 08/10/2010.
References

11. See http://web.chemistry.gatech.edu/~fernandez/DART_dataset.mat.


Downloaded from cebp.aacrjournals.org on June 13, 2017. © 2010 American Association for Cancer Research.
Cancer Epidemiology, Biomarkers & Prevention

Rapid Mass Spectrometric Metabolic Profiling of Blood Sera Detects Ovarian Cancer with High Accuracy


Cancer Epidemiol Biomarkers Prev  Published OnlineFirst August 10, 2010.

Updated version Access the most recent version of this article at: doi:10.1158/1055-9965.EPI-10-0126

Supplementary Material Access the most recent supplemental material at: http://cebp.aacrjournals.org/content/suppl/2010/08/09/1055-9965.EPI-10-0126.DC1

E-mail alerts Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.