A NOVEL PROFILE BIOMARKER DIAGNOSIS FOR MASS SPECTRAL PROTEOMICS

HENRY HAN[†1, 2]

[1]*Department of Computer and Information Science, Fordham University, New York NY 10023 USA* [2]*Quantitative Proteomics Center, Columbia University, New York 10027 USA*

Email: xhan9@fordham.edu

Mass spectrometry based proteomics technologies have allowed for a great progress in identifying disease biomarkers for clinical diagnosis and prognosis. However, they face acute challenges from a data reproducibility standpoint, in that no two independent studies have been found to produce the same proteomic patterns. Such reproducibility issues cause the identified biomarker patterns to lose repeatability and prevent real clinical usage. In this work, we propose a profile biomarker approach to overcome this problem from a machine-learning viewpoint by developing a novel derivative component analysis (DCA). As an implicit feature selection algorithm, derivative component analysis enables the separation of true signals from red herrings by capturing subtle data behaviors and removing system noises from a proteomic profile. We further demonstrate its advantages in disease diagnosis by viewing input data as a profile biomarker. The results from our profile biomarker diagnosis suggest an effective solution to overcoming proteomics data's reproducibility problem, present an alternative method for biomarker discovery in proteomics, and provide a good candidate for clinical proteomic diagnosis.

1. Introduction

With the recent surge in proteomics, large volumes of mass spectral serum/plasma/urine proteomic data are available to conduct molecular diagnosis in complex diseases. As a promising way to revolutionize medicine, mass spectral proteomics demonstrates a great potential in identifying novel biomarker patterns from a proteome for diagnosis, prognosis, and other diverse clinical needs [1,2]. However, robust clinical diagnosis from mass spectral data remains an acute challenge in translational bioinformatics due to the special characteristics of proteomics data.

First, mass spectral proteomics data are high-dimensional data that can be represented as a matrix $X \in \Re^{n \times p}$ after preprocessing, where each row represents protein expression at a mass-to-charge (*m/z*) ratio of peptides or proteins, usually called a feature from a machine learning perspective, and each column represents protein expression from a sample (observation) (e.g., a control or cancer subject) across all *m/z* ratios in an experiment. The number of rows is much greater than the number of columns, $p \ll n$, that is, #variables (peptides/proteins) is much greater than #samples. Usually $n \sim O(10^4)$ and $p \sim O(10^2)$. While there are a large amount of *m/z* ratios (peptides or proteins), only a few number of variables (e.g., peaks) have meaningful contribution to data variations and disease diagnosis. Moreover, they are not noise-free data because preprocessing and normalization methods themselves cannot remove built-in system noise from mass spectrometry technology itself. In fact, it remains a challenge to separate true signals in a mass spectral profile from red herrings though different endeavors from machine learning.

Second, mass spectral proteomics data usually suffer from data reproducibility problems, which mean that no two independent studies have been found to produce same proteomic patterns [2,3]. As such, corresponding biomarker patterns identified, which consists a small set of meaningful peaks, from these data may lose repeatability due to the poor reproducibility and

difficulty in validating biomarker patterns identified from multiple data sources. In fact, there are almost no reproducible biomarker patterns reported for mass spectral proteomic data in the literature [2]. Although several methods are proposed to mitigate this problem from a quantification perspective [2,3], there is no method to tackle this problem from a machine learning viewpoint as of yet.

The non-reproducibility of proteomic source data and their biomarker patterns, which are usually obtained by peak-selection methods using different machine learning algorithms, is mainly due to mass spectrometry technology's exquisite sensitivity to any subtle change in the proteome caused by biological or technical factors [3]. In other words, tiny changes in the proteome may lead to a set of completely different mass spectral peak patterns. Thus, a desirable diagnosis from identified protein or peptide biomarkers may not be reusable for other "same data" generated using the identical patient and control samples under the same profiling technologies and protocols.

In this work, we propose a *de novo* profile biomarker approach to achieve clinical level diagnosis. Unlike traditional biomarker discoveries that collect a few meaningful peaks, our profile biomarker approach views input data as a "whole biomarker" by proposing a novel derivative component analysis (DCA), which evolves from our previous work [4-6], and combing it with state-of-the-art classifiers. It is noted that a profile biomarker has the same dimension as the input data but with less variance and storage. In our approach, we aim at the reproducibility of diagnosis performance instead of looking for specific peptides or proteins, i.e., we believe a profile biomarker would be more robust than traditional biomarkers, provided it could achieve clinical level diagnoses for different proteomic data. That is, the motivation of this study is to solve the data reproducibility problem in proteomics by developing a novel profile biomarker diagnosis.

Our profile biomarker approach relies on a novel feature selection algorithm: derivative component analysis (DCA) as proposed in this work. Traditional feature selection algorithms (e.g., *t-test*) are usually characterized by the explicit feature number decrease or dimension reduction of the input data. It is noted that a feature refers to a row of protein/peptide expression of all samples at an *m/z* ratio. However, as an implicit feature selection algorithm, DCA conducts feature selection implicitly, i.e., there is no feature number decrease after DCA. More importantly, DCA enables the retrieval of the true signals from input proteomic data by removing redundant information and built-in noises, which provide a robust information support for our profile-biomarker diagnosis. Considering similar diagnosis mechanisms for proteomic profiles, we use benchmark serum proteomic data to demonstrate our profile biomarker diagnosis in this study.

The paper is organized as follows. Section 2 discusses essential components in profile biomarker discovery and proposes DCA in addition to addressing the weaknesses of the traditional feature selection methods. Section 3 investigates DCA-based profile biomarker diagnosis by integrating it with state-of-the-art classifiers. We further demonstrate our approach's superiority by comparing it with other state-of-the-arts, besides addressing DCA-induced biomarker discovery. Finally we discuss the pros and cons of our profile biomarker diagnosis and conclude our paper.

2. Derivative Component Analysis (DCA)

Before we proceed, we need to answer the question: 'what essential components are needed to make a profile biomarker successful in proteomics?' We believe that essential components for a profile biomarker approach may rely on whether we can separate true signals from red herrings for

each proteomic profile. Traditional feature selection methods usually fail to capture true signals from mass spectral proteomic data set because of their built-in weaknesses. Although various feature selection methods are employed in proteomics to glean informative features for the sake of diagnosis [7], there is no study to address their weaknesses systematically.

We categorize feature selection into input-space and subspace methods. The former seeks a feature subset $X' \in \Re^{m \times p}$, $m \ll n$, in the same space $\Re^{n \times p}$ as input data X by conducting a hypothesis test (e.g., *t-test*), or wrapping a classifier to select features recursively; the latter conducts a dimension reduction by transforming data into a subspace S induced by a linear or nonlinear transformation $f: X \to S$, where $S = span(s_1, s_{2,\ldots} s_k)$, $s_k \in \Re^k$, $k \leq p \ll n$, and seeking meaningful linear combinations of the features. For example, the subspace S will be spanned by all principal components when the transformation is induced by principal component analysis (PCA) [8]. In fact, almost all PCA, ICA, PLS, and NMF and their extensions such as nonnegative principal component analysis (NPCA), sparse NMF, and other related methods fall into this category [4-6,9]. However, the two types of methods have the following built-in limitations.

The weakness of the input and subspace methods. The input-space methods usually assume input data are clean or nearly clean, and lack de-noising schemes. The clean data assumption appears to be inappropriate for proteomic profiles because they usually contain nonlinear noise from technical or biological artifacts (e.g., built-in noise generated from profiling systems). The noise would enter feature selection as outliers and cause those peaks with less biological meaning to be selected, leading to an inaccurate or even poor decision function in classification and affecting the disease diagnosis and generalization.

On the other hand, subspace methods have difficulties capturing subtle data characteristics, because the subspace methods transform data into another subspace in order to seek meaningful feature combinations and the original spatial coordinates are lost in the transformation, which makes it almost impossible to track the mapping relationships between features and the specific data characteristics they interpret or contribute to. It is noted that subtle data characteristics refer to latent data behaviors interpreting transient data changes in a short time interval.

In contrast, global data characteristics refer to the holistic data behaviors interpreting long-time interval data changes, which happen more often than subtle data behaviors. The global data characteristics are easily extracted by general subspace methods like PCA, because there are more features contributing to holistic data behaviors than those contributing to subtle data behaviors. Furthermore, since most subspace methods treat all features uniformly regardless of which types of data behaviors they interpret, global characteristics are more likely to be selected than subtle data characteristics, because the former's features are more frequent than those of the latter in the feature domain.

As such, global data characteristics are usually over-extracted and subtle data characteristics may be totally missed or overshadowed after feature selection. The signals extracted from such feature selection are far from 'true signals' because the global data characteristics are over-expressed. The redundant global data characteristics would lead to a biased decision function for the following classifier (e.g., SVM) that favors the extracted global data characteristics, which may present a hurdle for clinical diagnosis, because the subtle characteristics are essential to achieve high-accuracy diagnosis for proteomics data, especially as different subtype tumor samples usually share similar or the same global data characteristics but different subtle data characteristics [5,6].

It is clear that the built-in weaknesses of the traditional feature selection methods prevent true

signal extraction and the possibility of profile biomarker diagnosis, because they lack de-noising and subtle data characteristics retrieval schemes. We sketch the key reasons for these weaknesses as follows before we present our derivative component analysis.

The reasons for traditional feature selection's weaknesses. The following are the major reasons why traditional feature selection methods are unable to extract subtle characteristics and remove systems noise effectively. 1) These methods are single resolution data analysis methods that view each feature as an indivisible information unit, which makes system noise removal almost impossible; 2) They treat all features uniformly regardless of their frequencies in the feature space, which makes subtle data characteristics extraction difficult due to lower frequencies in the feature domain. Mathematically, retrieving subtle data characteristics, which are represented by transient data behaviors, means to seek the derivative of the original data. However, this is theoretically quite difficult to complete in a single resolution mode.

Derivative component analysis (DCA). We propose a novel feature selection algorithm: derivative component analysis (DCA) to separate true signals from red herrings, that is, conduct de-noising for system noise and retrieve subtle data characteristics in a multi-resolution data analysis mode. As a multi-resolution feature selection algorithm, the proposed DCA no longer views a feature as an indivisible information element. Instead, all features are hierarchically decomposed into different components to discover data derivatives so as to capture the subtle data characteristics and conduct de-noising. The proposed derivative component analysis (DCA) mainly consists of the following three steps.

First, a discrete wavelet transform (DWT) [10] is applied to all features to decompose it hierarchically as a set of detail coefficient matrices $cD_1, cD_2 \cdots cD_J$ and an approximation matrix cA_J under a transform level *J*. It is worthwhile to point out that we view each *m/z* ratio as a corresponding time point in our context for the convenience of the DWT [10]. Since the DWT is calculated on a set of dyadic grid points hierarchically, the dimensionalities of the approximation and detail coefficient matrices shrink dyadically level by level.

It is noted that the approximation matrix and coarse level detail coefficient matrices (e.g. cD_J) capture global data characteristics, because they contain contributions from those features contributing to data behaviors in 'long-time windows', and outlining the global tendency of the data. Similarly, the fine level detail coefficient matrices (e.g., cD_1, cD_2) capture subtle data characteristics, because they contain contributions from those features that disclose quick changes in 'short-time windows', and describe data derivatives locally. In fact, these fine level detail matrices are the components for reflecting the data derivatives in different short-time windows. As such, they can be called 'derivative components' for the functionality in describing data behaviors.

Furthermore, most system noises are transformed in these derivative components due to its heterogeneity with respect to the features contributing to the global tendency of data. Clearly, the DWT in the first step separates the global characteristics, subtle data characteristics, and noises in different resolutions.

Second, retrieve the most important subtle data characteristics and conduct de-noising by reconstructing these fine level detail coefficient matrices before or at a presetting cutoff level τ (e.g., τ=*3*). Such a construction is summarized in two steps: 1) Conduct principal component analysis (PCA) for the detail matrices $cD_1, cD_2 \cdots cD_\tau$. 2) Reconstruct each detail coefficient matrix by using its first *m* principal components, in each principal component (PC) matrix. Usually, $m = 1$, i.e., we employ the first PC to reconstruct each detail coefficient matrix, which means we only retrieve the most important subtle data characteristics in the detail coefficient matrix

reconstruction. In fact, the first PC based reconstruction also achieves de-noising by suppressing the noises' contribution in the detail coefficient matrix reconstruction because the noises are usually least likely to appear in the 1st PC.

On the other hand, those coarse level detail coefficient matrices after the cutoff τ: $cD_{\tau+1}, cD_{\tau+2} \cdots cD_J$ and approximation coefficient matrix cA_J are kept intact to retrieve global data characteristics. In fact, the parameter m can be also determined by using a variability explanation ratio ρ_m defined as follows, such that it is greater than a threshold ρ (e.g., $\rho = 60\%$), which is the variability explanation ratio interpreted by the first principal components of the detail coefficient matrices before or at the cutoff.

Variability explanation ratio. Given a data set with n variables and p observations, usually, $p<n$, the variability explanation ratio is the ratio $\rho_m = \sum_{i=1}^{m} \sigma_i / \sum_{i=1}^{p} \sigma_i$ between the variance explained by the first m PCs and the total variances, where σ_j is the variance explained by the j^{th} PC, which is the j^{th} eigenvalue of the covariance matrix of the input proteomic data.

Such a selective reconstruction process extracts the most important subtle data characteristics and achieves de-noising by suppressing the noises' contribution to the fine detail coefficient matrix reconstruction. This is because only the 1st PC or few top PCs are employed to reconstruct each targeted fine level coefficient matrix cD_j and the other less important and noise-contained principal components are dropped in reconstruction.

Third, conduct the corresponding inverse DWT by using the current detail and approximation coefficient matrices to obtain meta-data X_*, which is a de-noised data set with subtle data characteristics extraction, because of the highlight of the most significant subtle data behaviors in the "derivative components" based reconstructions. The meta-data are just 'true signals' separated from red herrings that share the same dimensionality with the original data but with less memory storage because less important PCs are dropped in our reconstruction.

It is noted that, unlike traditional feature selection methods, DCA is an implicit feature selection method, where useful characteristics are selected implicitly without an obvious variable removal or dimension reduction. Algorithm 1 gives the details about DCA as follows, where we use X^T instead of X for the convenience of description, i.e., each row is a sample and each column is a feature.

Algorithm 1 Derivative Component Analysis (DCA)

1. **Input:** $X^T = [x_1, x_2, \ldots x_n]$, $x_i \in \Re^p$, *DWT level J*; *cutoff* τ; *wavelet* ψ; *thereshold* ρ;
2. **Output:** Meta-data X_*^T
3. **Step 1.** *Column-wise discrete wavelet transforms (DWT)*
4. Conduct J-level DWT with wavelet ψ for each column of X^T to obtain
5. $[cD_1, cD_2 \ldots cD_J; cA_J]$, $cD_j \in \Re^{p_j \times n}$, $cA_J \in \Re^{p_J \times n}$, and $p_j = \lceil p / 2^j \rceil$, $j = 1,2,\ldots J$.
6. **Step 2.** *Subtle data characteristics extraction and de-noising*
7. *for* j = 1 to J
8. *if* $j \le \tau$
9. a) Do principal component analysis for each detail matrix cD_j to obtain its PC and score matrix
10. $U = [u_1, u_2, \ldots u_{p_j}]$, $u_i \in \Re^n$ and $S = [s_1, s_2, \ldots s_{p_j}]$, $s_i \in \Re^{p_j}$, $i = 1,2,\cdots, p_j$.
11. b) Reconstruct matrix cD_j by employing first m principal components $u_1, u_2, \ldots u_m$, *s.t.* $\rho_m \ge p$
12. $cD_j \leftarrow cD_j \times (I \times I^T) / p_j + \sum_{i=1}^{m} u_i \times s_i^T$, $I = [1,1,\cdots 1]^T \in \Re^{p_j}$

13. *end if*
14. *end for*
15. **Step 3.** Approximate the original data by the inverse discrete wavelet transform
16. $X_*^T \leftarrow inverseDWT([cD_1, cD_2 \ldots cD_J; cA_J])$ with the wavelet ψ

Although an optimal DWT level can be obtained theoretically according to the maximum entropy principle [11], it is reasonable to adaptively select the DWT level J according to the *'nature'* of input data, where large #samples corresponds to a relatively large J value, for the convenience of computation. As such, we select the DWT level as $4 \leq J \leq \lceil \log_2 p \rceil$ considering the magnitude level of the samples number p in proteomics data to avoid too large or too small transform levels. Correspondingly, we empirically set the cutoff as $1 < \tau \leq J/2$ to separate the fine and coarse level detail coefficient matrices for good performance.

Furthermore, we require the wavelet ψ in the DWT orthogonal and have compact supports such as *Daubechies* wavelets (e.g., '*db8*'), for the sake of subtle data behavior capturing. Interestingly, we have found that the first PC of each fine level detail coefficient matrix usually has a quite high variability explanation ratio (e.g., >60%) for each fine level detail coefficient matrix cD_j ($1 \leq j \leq \tau$). Thus, we relax the variability explanation ratio threshold by only using the first PC to reconstruct each cD_j matrix in order to catch subtle data characteristics along the maximum variance direction. In fact, we have found that using more PCs in the fine level detail coefficient matrix reconstruction does not demonstrate advantages in subtle data characteristics extraction and de-noising than using the first PC.

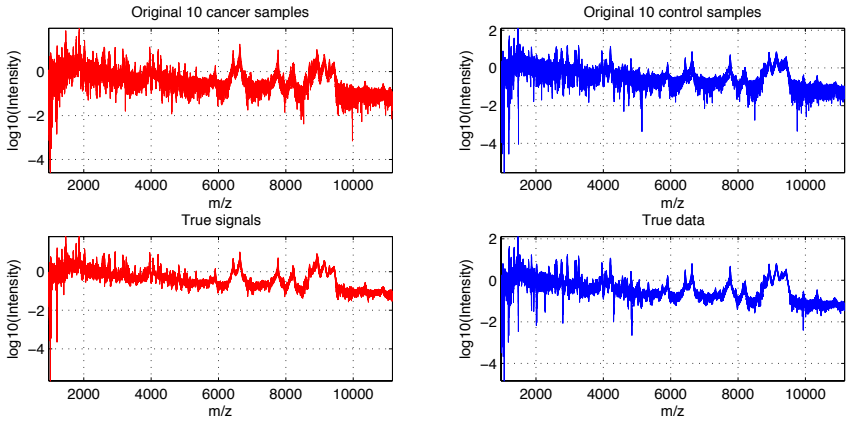

Fig 1. The true signals of 10 cancer and control samples across 16331 *m/z* of the *Colorectal* data by DCA

Figure 1 shows the true signals (meta-data) of the 10 cancer and control samples, which are randomly selected from *Colorectal* data [12] with total 48 controls and 64 cancer samples across 16331 *m/z* ratios, extracted by our DCA under the cutoff τ=*2*, transform-level J = *7*, and wavelet *'db8'*. Interestingly, the each type of samples in the extracted true signals appear to be smoother and more proximal to each other besides demonstrating less variations, because of the major subtle data characteristics extraction and system noise removal.

Such a case is demonstrated more clearly by Figure 2, where the 10 cancer and control samples and their true signals are highlighted between 1400 Da and 1500 Da. It is quite clear to observe that the same type samples are closer to each other spatially, and some small spikes are removed as the built-in noises in true signals. Obviously, from a classification viewpoint, these

true signals will contribute to high accuracy diagnoses than the original proteomic data, because the built-in noises and redundant global data characteristics would have a much lower chance to get involved in classification due to derivative component analysis. Instead, subtle data characteristics would have a greater chance of participating in the decision rule inference.

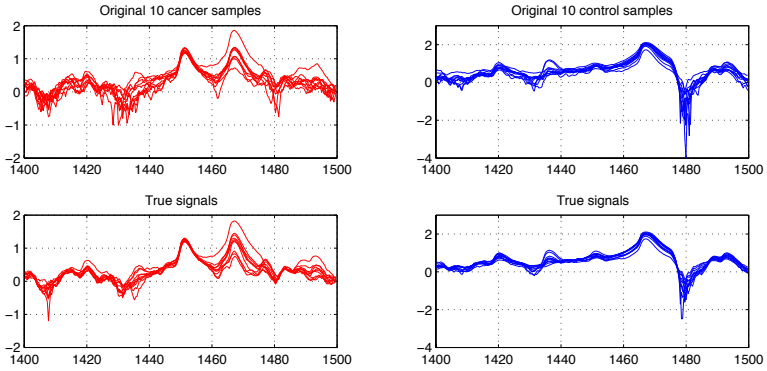

Fig 2. The true signals of 10 cancer and control samples of the *Colorectal* data between 1400-1500 Da.

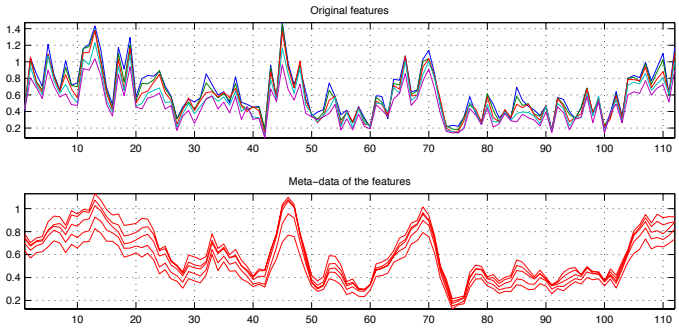

Fig 3. Random five features in *Colorectal* data and its meta-data across 112 samples (64 cancers + 48 controls).

Similarly, Figure 3 shows the meta-data of randomly picked five features from *Colorectal* data under the same parametric setting for DCA. Interestingly, the meta-data (meta-features) are *smoother* and have *values in a smaller range* than the original features for its subtle data characteristics extraction and de-noising. The meta-features are actually more distinguishable than their original features, which reflect the true expression level of the peptides/proteins at the *m/z* ratios better. In other words, DCA provides a 'zoom' mechanism to capture the original data's subtle behaviors that are usually latent in general machine-learning methods.

Profile Biomarker Diagnosis with DCA

Since DCA can separate true signals from red herrings by extracting subtle data characteristics and removing built-in noises, it is natural to combine DCA with start-of-the-art classifiers to conduct profile biomarker diagnosis, where input proteomics data are viewed as a profile biomarker. We chose support vector machines (SVM) for its efficiency and advantages in handling large-scale data, popularity in proteomics diagnosis and biomarker discovery [13]. As such, we propose novel derivative component analysis-based support vector machines (DCA-

SVM) in order to attain a profile biomarker disease diagnosis, which is actually equivalent to a binary or multi-class classification problem.

Given a binary type training samples $X=[x_1,x_2,\cdots x_p]^T$ and their labels $\{x_i,c_i\}_{i=1}^p$, $c_i\in\{-1,1\}$, its corresponding meta-data $Y=[y_1,y_2,\cdots y_p]^T$ are computed by using DCA, Then, a maximum-margin hyperplane $O_h: w^T y+b=0$ in $\Re^n$ is constructed to separate the '+1' ('cancer') and '-1' ('control') types of the samples in the meta-data Y, where w and b are the normal and offset vector of the hyperplane respectively. The hyperplane construction is equivalent to solving the following quadratic programming problem (standard SVM, i.e., C-SVM):

$$\begin{aligned} &\min_{w,b,\xi} \tfrac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{p}\xi_i \\ &s..t.\ c_i(w^T y_i+b)\geq 1-\xi_i, i=1,2\ldots p \\ &\xi_i\geq 0 \end{aligned} \tag{1}$$

The C-SVM can be solved by seeking the solutions to the variables α_i of a corresponding Lagrangian dual problem to get a decision function $f(x')=sign(\sum_{i=1}^{n}\alpha_i c_i k(y_i\bullet y')+b)$ to determine the class type of a testing sample x', where y' and y_i are corresponding meta-samples computed from DCA for samples x' and x_i. The kernel function $k(y_i,y)$ maps y_i and y' into a same-dimensional or high-dimensional feature space. In this work, we employ the *'linear'* kernel for its simplicity and efficiency. Our multiclass DCA-SVM algorithm employs the *'one-against-one'* to conduct multiclass phenotype diagnosis for its proved advantage over the *'one-against-all'* and *'directed acyclic SVM'* methods [14].

It is worthwhile to point out that our DCA-SVM has a different feature space due to true signal extraction from DCA. The standard SVM's feature-space usually contains noises from input proteomic data, and misses subtle data characteristics. Alternatively, the DCA-SVM 's feature space contains 'de-noised' true signals with subtle data characteristics, which avoids the global data characteristics favored decision rule because subtle data characteristics are also invited in SVM hyperplane construction besides the global data characteristics. As such, the DCA-SVM can efficiently detect those samples with similar global characteristics but different subtle characteristics in disease diagnosis than the standard SVM.

3. Results

We demonstrate our profile biomarker diagnosis' superiority by using five benchmark serum proteomic data sets, which include *Cirrhosis, Colorectal, HCC, Ovarian-qaqc* and *ToxPath* data [12,15-17,19]. The benchmark data used in our experiments are heterogeneous data generated from different experiments via different profiling technologies such as MALDI-TOF and SELDI-TOF, and preprocessed by different methods. Table 1 describes the details of the five data sets.

We compare the proposed DCA-SVM based profile-biomarker diagnosis with the following state-of-the-arts in this work. They include a partial least square (PLS) based linear logistic discriminant analysis (PLS-LLD) [18,20], standard SVM [13], a SVM combining with principal component analysis: PCA-SVM [5], and a SVM with input-space feature selection: *fs*-SVM, which employs *t-test* and *Anona1 (one-way ANOVA)* to conduct feature selection for binary and multi-class data respectively. For each data, *fs*-SVM collects a meaningful feature set including all

features with *p-values* < *0.05* using *t-test* or *Anova1* for phenotype diagnosis.

Table 1. Benchmark proteomic data

Data	#Feature	#Sample	Platform
Cirrhosis	23846	72 controls + 78 HCCs + 51 cirrhosis	MALDI-TOF
Colorectal	16331	48 controls + 64 cancers	MALDI-TOF
HCC	6107	181 controls +176 cancers	SELDI-QqTOF
Ovarian-qaqc	15000	95 controls + 121 cancers	SELDI-TOF
ToxPath	7105	28 normals + 43 potential normals + 34 cardiotoxicities + 10 potential cardiotoxicities	SELDI-QqTOF

We employ the *'linear'* kernel $k(x, y) = (x \cdot y)$ in all SVM-related classifiers for its efficiency in omics data classification, rather than nonlinear kernels (e.g., Gaussian kernels), which usually lead to overfitting in diagnosis [4-6]. To avoid potential biases from presetting training/test data partition on diagnosis, we employ the k-fold (k=5) cross-validation to evaluate the five classifiers' performances for all data sets. In addition to choosing the first ten PLS components in the PLS-LLD classifier, we uniformly set the DWT level $J = 7$ under '*db8*', cutoff $\tau = 2$; and apply the first PC-based detail coefficient matrix reconstruction in DCA to retrieve true signals for all proteomic data sets.

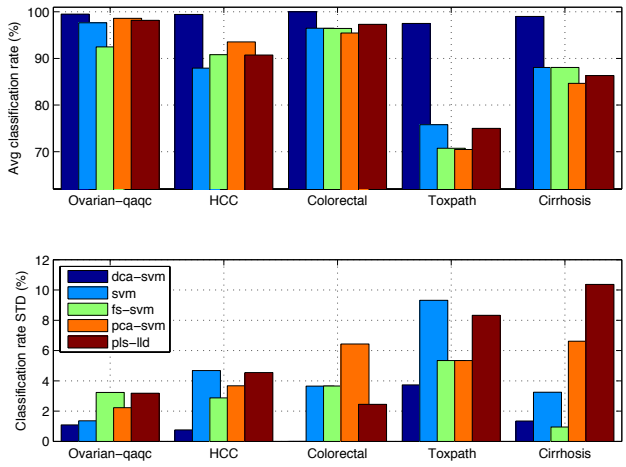

Fig 4 Comparing profile biomarker diagnosis' diagnostic accuracies and its standard deviations with those of others.
s

Before demonstrating our profile biomarker approach's advantages, we introduce several key diagnosis performance measures, namely, diagnostic accuracy, sensitivity, specificity and positive predication ratios, as follows. The diagnostic accuracy is the ratio of the correctly classified test samples over total test samples. The sensitivity, specificity, and positive predication ratio are defined as the ratios: $\frac{TP}{TP+FN}$, $\frac{TN}{TN+FP}$, and $\frac{TP}{FP+TP}$ respectively, where *TP(TN)* is the number of positive (negative) targets (a positive (negative) target is a proteomic sample with '+1' ('-1') label) correctly diagnosed and *FP (FN)* is the number of negative (positive) targets incorrectly

diagnosed by the classifier.

Figure 4 demonstrates rivaling clinical level performance from our profile biomarker diagnosis (DCA-SVM) by comparison with the other classifiers in average diagnosis accuracies and its standard deviations. It seems that our profile biomarker diagnosis achieves performance nearly clinical level and demonstrate strongly leading advantages over its peers in a stable manner. Alternatively, those comparison classifiers seem to show quite large level oscillations that may indicate they lack stability and good generalization capacities across different data sets, which exclude themselves as candidates for clinical proteomics diagnosis.

For example, our profile biomarker diagnosis achieves 99.52% (sensitivity 100%, specificity 99.17%), 100% (sensitivity 100%, specificity 100%), and 99.44% (sensitivity 98.00%, specificity 100%) diagnostic accuracies on the *Ovarian-qaqc*, *Colorectal* and *HCC* data respectively. It further reaches 97.50%, 99.01% diagnostic accuracies for *Toxpath* and *Cirrhosis* data respectively. However, the standard SVM classifier can only achieve 75.80% and 88.06% diagnosis for the same data sets respectively. Although some input-space or subspace methods may sometimes boost diagnosis for binary-type data set, we have found that they are unable to increase the SVM classifier's diagnosis and generation abilities significantly, especially for multiclass proteomic data. In fact, in contrast to the proposed profile biomarker diagnosis, all the comparison classifiers show high-level oscillations in diagnoses across different data sets. It is noteworthy that the high-level oscillations in diagnosis is further highlighted by corresponding large standard deviation values in diagnosis from those classifiers in Figure 4, where our DCA-SVM based profile biomarker diagnosis demonstrates its good stability and generalization for its smallest standard deviation values across all the data sets.

Compare profile biomarker diagnosis with prior methods. It is worthwhile to point out that our DCA-SVM based profile biomarker also demonstrates its superiority to its peers in terms of diagnostic accuracy, sensitivity, specificity and positive predication ratios. We further compare our profile biomarker diagnosis approach with the previous biomarker discovery diagnoses in the literature and have found that our method demonstrates good clinical level sensitivities in phenotype discriminations for different benchmark proteomic data. For example, Alexandrov et al 's work only achieved 97.5% diagnosis accuracy with sensitivity 98.4% and specificity 95.8% for *Colorecta* data by using a complicated method [12]. However, our profile biomarker diagnosis achieves 100% diagnosis accuracy with sensitivity 100% and specificity 100%. For *Ovarian-qaqc* data, our approach achieves a 99.53% clinical-level diagnosis accuracy with sensitivity 98.95% and specificity 100%, which is better than the original diagnosis level obtained in [17] and all the other peers. For *Cirrhosis* data, Ressom *et al* partitioned this three-class data into two binary data sets and proposed a novel hybrid ant colony optimization based support vector machines (ACO-SVM), where ACO was used for biomarker discovery, to achieve 94% and 100% specificity to distinguish hepatocellular carcinoma (HCC) from cirrhosis [16]. There was no result available to distinguish normal, HCC, and cirrhosis in a multiclass diagnostic way. However, our proposed approach has achieved 99.01% diagnosis accuracy for this multi-class data set.

Can DCA be used to conduct biomarker discovery by collecting meaningful peaks if we relax the reproducibility concern? The answer is 'yes' because derivative component analysis can identify meaningful protein or peptide peaks from true signals We simply apply *t-test* and *Anova1* to identify the top-ranked features with the smallest *p-values*, i.e. we pick the three top-scored peaks as biomarkers for its statistical significance. Figure 5 illustrates the separation of four benchmark data sets with three top-ranked biomarkers (peaks). It is interesting to see that these

high-dimensional proteomic profiles can be separated almost completely with these biomarkers identified from true signals.

We can also obtain some meaningful biological depth by checking these biomarkers. For example, the SW plot in Figure 5 shows the separation of 176 controls and 181 cancers in the *HCC* data, by the top-ranked biomarkers (peaks) at 2534.2, 2584.3, and 6486.2 *m/z* ratios, where each dot represents a sample (a patient with HCC or a healthy subject). It is also interesting to see that two biomarkers are from downsteam *m/z* ratios, which were believed to be more sensitive to detect phenotype information than those from upstream *m/z* ratios [16,19]. Moreover, The separation can provide meaningful biological insight for pathological disease states. For example, we select three top-ranked biomarkers at 1668.99, 5907.73, 5907.13 *m/z* ratios for the *Cirrhosis* dataset, which is a three-class high-resolution MALDI-TOF proteomic profile with 23846 features. The phenotype separations provided by the three biomarkers give very meaningful biological insights, i.e., the SE plot in Figure 5 shows the three clearly independent clusters, where Cirrhosis cluster with 51 samples (blue) have closer spatial distances to the HCC cluster 78 samples (red) than the normal cluster with 72 samples (yellow). Such spatial distances demonstrated by our biomarkers are actually consistent to their pathological distances: Cirrhosis is the middle stage to hepatocellular carcinoma (HCC) for a healthy subject.

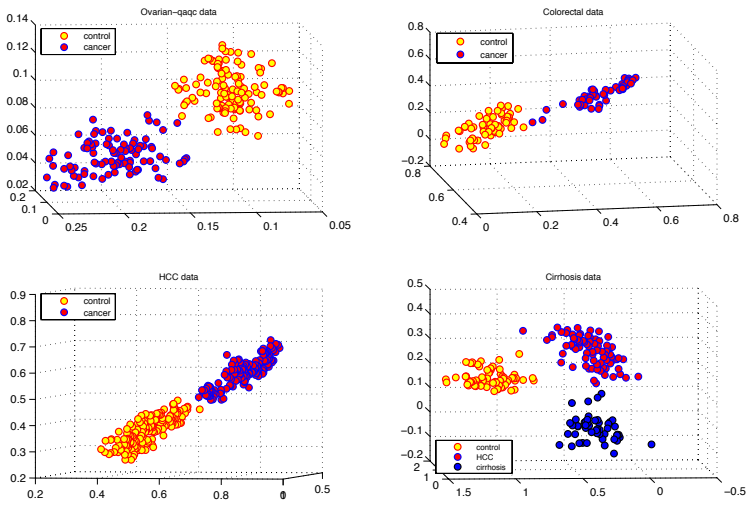

Fig 5 Separating disease phenotypes of four data sets by only using their three biomarkers with the smallest p-values.

4. Conclusions and Discussion

In this study, we propose a profile biomarker diagnosis approach to overcome the data reproducibility issue in proteomics data and demonstrate its clinical level performances across different data. The profile biomarker diagnosis is based on the novel implicit feature selection algorithm: derivative component analysis and derivative component analysis based support vector machines proposed in this study. As an implicit feature selection algorithm, DCA is able to separate true signals from red herrings by extracting subtle data characteristics and removing system noise via calculating a same dimensional meta-data for input proteomic data. It is noted that the complexity of DCA is higher than that of PCA, because DCA calls the classic PCA in several fine level detail coefficient matrix reconstruction, in addition to the DWT and inverse

DWT. However, DCA demonstrates a promising way to overcome the data reproducibility issue in proteomics because the high-accuracy diagnosis results seem to be reproducible themselves for different data sets under our approach. In other words, our profile biomarker diagnosis presents itself as an ideal candidate to achieve clinical diagnosis in clinical proteomics. Furthermore, our work suggests a key issue in proteomic disease diagnosis, that is, subtle data characteristics gleaning and de-noising can be more important in proteomics data feature selection and following phenotype discrimination than dimension reduction. Moreover, the proposed derivative component analysis provides an alternative feature selection by implicitly extracting useful data characteristics whiling maintaining the data's original dimensionality.

Although we are quite optimistic to see that our profile biomarker diagnosis will be a potential candidate to achieve a clinical disease diagnosis in proteomics by conquering the reproducibility problem, rigorous proteomics clinical tests are needed urgently to explore such a potential and validate its clinical effectiveness. In our ongoing work, we are working with pathologists to investigate extending the profile biomarker diagnosis approach to TCGA and RNA-Seq data besides protein expression array analysis.

5. References

1. T. Rath et al, Serum Proteome Profiling Identifies Novel and Powerful Markers of Cystic Fibrosis Liver Disease, *PLoS ONE*, (2013)
2. J. Ioannidis et al, Improving Validation Practices in "Omics" Research, *Science* **334**, 1230, (2011)
3. R. Hüttenhain et al, Reproducible Quantification of Cancer-Associated Proteins in Body Fluids Using Targeted Proteomics, *Sci Transl Med* **4**, 142ra94, (2012)
4. X. Han, Nonnegative Principal component Analysis for Cancer Molecular Pattern Discovery, IEEE/ACM Transaction of Computational Biology and Bioinformatics **7** (3), p537-549, (2010)
5. X. Han, Nonnegative principal component analysis for mass spectral serum profiles and biomarker discovery, *BMC Bioinformatics*, **11**(Suppl 1): S1, (2010)
6. H. Han, and X. Li, Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery, *BMC Bioinformatics*, 12(S1):S7, (2011)
7. M. Hilario and A. Kalousis, Approaches to dimensionality reduction in proteomic biomarker studies, *briefings in bioinformatics*, **9**:2 101-119, (008)
8. I. Jolliffe, Principal component analysis, Springer, New York, (2002)
9. J. Brunet, et al, Molecular pattern discovery using matrix factorization, *PNAS* **101**(12),4164–69, (2004)
10. S. Mallat, A wavelet tour of signal processing, Acad. Press, CA, USA, (1999)
11. T. Kapur and A. Kesevan, *Entropy optimization principles with applications*, Academic Press, (1992)
12. T. Alexandrov et al, Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation, Bioinformatics, Vol. 25(5):643-649, (2009)
13. V. Vapnik, *Statistical Learning Theory*. John Wiley, New York, (1998)
14. C. Hus and C. Lin, A Comparison of Methods for Multi-class Support Vector Machines, *IEEE Transactions on Neural Networks*, **13** (2):415-425, (2002)
15. H. Ressom et al, Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics* **21**(21), 4039-4045, (2005)
16. H. Ressom et al, Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* **23**(5), 619-626, (2007)
17. T. Conrads et al, High-resolution serum proteomic features for ovarian detection, *Endocrine-Related Cancer*, **11**, 163-178 (2004)
18. D. Nguyen, and D. Rocke, Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics* **18**:39–50, (2002)
19. E. Petricoin et al, Toxicoproteomics: serum proteomic pattern diagnostics for early detection of drug induced, *Toxicologic Pathology*, **32** (Suppl. 1):1–9, (2004)
20. D. Sampson et al A Comparison of Methods for Classifying Clinical Samples Based on Proteomics Data: A Case Study for Statistical and Machine Learning Approaches, *PLoS One*, (2011)