BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 15 2006, pages 1902–1909
doi:10.1093/bioinformatics/btl276

Data and text mining

A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS

Matthew Bellew[1], Marc Coram[2], Matthew Fitzgibbon[2], Mark Igra[1,2], Tim Randolph[2], Pei Wang[2], Damon May[2], Jimmy Eng[2], Ruihua Fang[2], ChenWei Lin[2], Jinzhi Chen[2,3], David Goodlett[3], Jeffrey Whiteaker[2], Amanda Paulovich[2] and Martin McIntosh[2,*]

[1]LabKey Software, Seattle, WA 98109, USA, [2]Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA and [3]University of Washington, Seattle, WA 98195, USA

Received on March 6, 2006; revised on May 26, 2006; accepted on May 29, 2006
Advance Access publication June 9, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Comparing two or more complex protein mixtures using liquid chromatography mass spectrometry (LC-MS) requires multiple analysis steps to locate and quantitate natural peptides within a single experiment and to align and normalize findings across multiple experiments.

Results: We describe msInspect, an open-source application comprising algorithms and visualization tools for the analysis of multiple LC-MS experimental measurements. The platform integrates novel algorithms for detecting signatures of natural peptides within a single LC-MS measurement and combines multiple experimental measurements into a peptide array, which may then be mined using analysis tools traditionally applied to genomic array analysis. The platform supports quantitation by both label-free and isotopic labeling approaches. The software implementation has been designed so that many key components may be easily replaced, making it useful as a workbench for integrating other novel algorithms developed by a growing research community.

Availability: The msInspect software is distributed freely under an Apache 2.0 license. The software as well as a Zip file with all peptide feature files and scripts needed to generate the tables and figures in this article are available at http://proteomics.fhcrc.org/

Contact: mmcintos@fhcrc.org

Supplementary Information: Supplementary materials are available at http://proteomics.fhcrc.org/CPAS (select 'Published Experiments' from the list of Projects and then 'msInspect Paper').

1 INTRODUCTION

Researchers hope to improve our understanding of biology as well as potentially revolutionize clinical care by identifying protein biological markers, or biomarkers, that can classify disease (Etzioni *et al*., 2003). Mass spectrometry has become a key technology for comparing biosamples because of its ability to broadly survey the peptide or protein constituents of the samples.

The use of mass spectrometry to identify protein biomarkers was first popularized using SELDI or MALDI technologies (Adam *et al*., 2002; Listgarten and Emili, 2005; Morris *et al*., 2005; Petricoin *et al*., 2002; Randolph and Yasui, 2006; Vlahou *et al*., 2001; Yasui *et al*., 2003), which locate biomarkers in MS data by their mass-to-charge ratio (*m/z*). Liquid chromatography mass spectrometry (LC-MS) provides a two-dimensional approach to profiling a proteome and allows researchers to locate peptides by their *m/z* and their LC retention time, both of which are related to the chemical structure of the peptides. The additional level of separation offered by LC-MS may have several advantages over one-dimensional approaches, including better sensitivity and resolution. With both platforms the sequences of the potential biomarkers are obtained by targeting their locations in a second interrogation using tandem MS (e.g. with LC-MS/MS or MALDI-MS/MS) with a compatible LC configuration (Domon and Aebersold, 2006).

In their comprehensive review, Listgarten and Emili (2005) point out that using this two-stage approach with LC-MS data is recent but not entirely new, and they describe several additional computational challenges that must be addressed to extend one-dimensional approaches into two dimensions. For example, unlike in MALDI, peptides in LC-MS can obtain multiple charges, which must be ascertained in order to compute a peptide's mass. Moreover, comparing peptide intensities across multiple experiments requires aligning in two dimensions rather than in one, and the additional retention time dimension varies unpredictably and non-linearly.

Many researchers have implemented algorithms to address one or more of these aspects of the problem, including the identification of peptides within a single experiment, quantitation of peptides and alignment across runs (Li *et al*., 2005). One of the earliest implementations is by Smith *et al*. (2002) who describe an approach to evaluate multiple LC-MS datasets and align them using a separate accurate mass tag database. A more recent open-source platform is mzMine, introduced by Katajamma *et al*. (Katajamaa *et al*., 2006; Katajamaa and Oresic, 2005), which also includes a graphical interface. Other researchers have recently begun to focus on improving specific components of the analysis process rather than providing complete comprehensive platforms. These components include the peculiarities of normalization across LC-MS runs (Callister *et al*., 2006; Wang *et al*., 2006b) and better techniques for matching peptides across multiple experiments (Wang *et al*., 2006a).

*To whom correspondence should be addressed.

We describe a software platform called msInspect (Mass Spectrometry *In Silico* Peptide Characterization Tool) that provides a comprehensive pipeline for quantitatively comparing peptides from biological samples using LC-MS data. msInspect processes individual LC-MS data files to locate peptides in two dimensions, and quantitates them by their signal intensity. Comparison of samples can be performed within an experiment using procedures for identifying isotopically labeled pairs (e.g., ICAT, 16O/18O, SILAC) or between experiments using a label-free approach that aligns and then normalizes multiple experiments. The label-free approach is intended to process multiple LC-MS experiments to form a peptide array, at which point standard methods for analyzing genomic arrays can be applied to classify the samples.

The algorithms in msInspect include multiple components specifically designed for LC-MS data. The signal processing component exploits the two-dimensional nature of the data to identify co-eluting isotopes and then groups them based on the similarity of the observed isotopic distributions to those of natural peptides. The alignment method estimates underlying non-linear mapping of retention times between experiments. The normalization approach adapts methods developed for genomic arrays to accommodate natural variation of LC-MS signal intensities across runs. msInspect was designed to serve as a workbench for disseminating and developing novel algorithms; many components of the signal processing, alignment and normalization procedures can be replaced without having to alter either the framework of other supporting algorithms or the downstream visualization tools. msInspect also supports interaction with the freely available R statistical programming language.

The ability to identify biomarkers will depend on the reproducibility of the analysis platform and also on the number and size of differences that exist between any two types of biological samples. This manuscript demonstrates msInpsect's algorithms using three sets of experiments. The first two establish the reproducibility of labeling and label-free approaches from independent interrogations of human serum. The third demonstrates the ability to use tools from genomic array analysis to classify mutant and wild-type forms of bacteria.

2 METHODS

2.1 Data inputs

The msInspect program accepts as input one or more LC-MS data files represented in the standard mzXML data format (Pedrioli *et al*., 2004) from instrumentation with resolution high enough to discern isotopic distributions.

2.2 Graphical interface

The msInspect graphical interface allows a user to view and interrogate LC-MS raw data and examine the results of the peptide location algorithms (Figure 1). The primary pane (1A) displays an image of the raw data from a single LC-MS measurement; the horizontal and vertical axes represent retention time and *m/z*, respectively, and the darkness of the color represents signal intensity. The red points in Figure 1A represent the monoisotopic mass and maximum intensity of each of the located peptides, the result of step 1 described below. The user can overlay the peptides found in other LC-MS data files as well (data not shown). Users can inspect a detailed, close-up view of any smaller region of the image. Selecting a location provides both a detailed two-dimensional image of the local region (Fig. 1B) as well as one-dimensional cross sections of the mass- or retention time-dimensions (Fig. 1C, detailed mass dimension shown here). Other

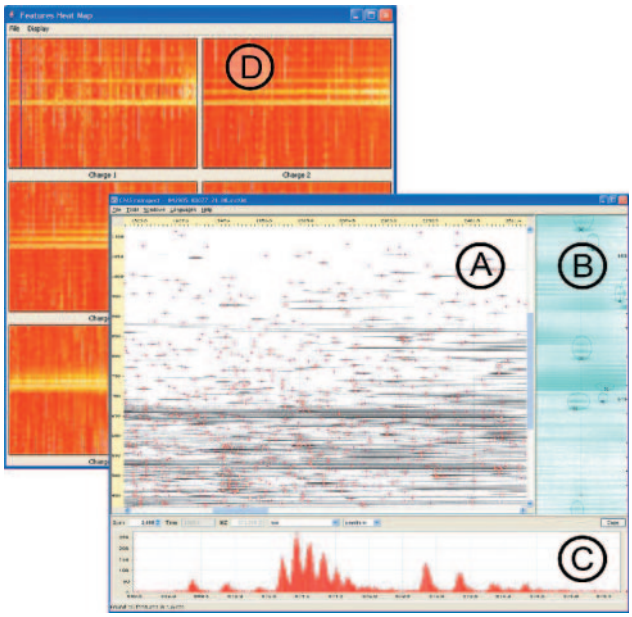

Fig. 1. The msInspect Graphical Interface. The main window displays: (**A**) an image of the mzXML file; (**B**) a tighter view of an area in the image that the user selects and (**C**) *m/z* spectra and elution profiles corresponding to the point in A that the user selects. The Heat Map tool (**D**) is used for visual curation.

graphical functions include the ability to zoom in and out of the pane in Figure 1A, to display the properties, such as intensity, of specific peptides and to visually curate the results (Fig. 1D and described below).

2.3 Workflow

All functions are accessible via interactive tools and also via command line. The functionality of msInspect can be described in three parts. Part 1 involves processing individual LC-MS images to produce a tab delimited peptide feature file listing the mass and retention time of all peptides and other descriptive information (see Section 3.1). Part 2 includes algorithms that operate on individual peptide feature files such as allowing visual curation and correction of the findings and providing further processing to identify isotopically labeled pairs and their intensity ratios. Part 3 supports label-free quantitation, including alignment of multiple peptide feature files into a single peptide array and normalization to accommodate natural variation across experiments. The peptide array may then be further evaluated using tools developed for analyzing genomic arrays. These three parts can be used together in a sequence or independently using the output from previous steps. More details can be found in Supplementary Material (Figure S1).

2.4 Software architecture

The msInspect software package is written in platform-independent Java with alignment and normalization routines implemented in the freely available R statistical programming language. Many msInspect components are modular so they can be replaced easily while maintaining the overall graphical features. For example, alignment and normalization routines may be easily replaced by exchanging R routines that have the defined inputs and outputs. New signal processing algorithms, described in Part 1, can be discovered at run-time and dynamically loaded by the Java class loader. In addition to the two-dimensional algorithm described in this article, several other signal processing implementations are provided for special purposes (e.g. finding peptide features within a single scan).

The msInspect program is available under an Apache 2.0 License via executables and via the Java Web Start package. In addition, msInspect

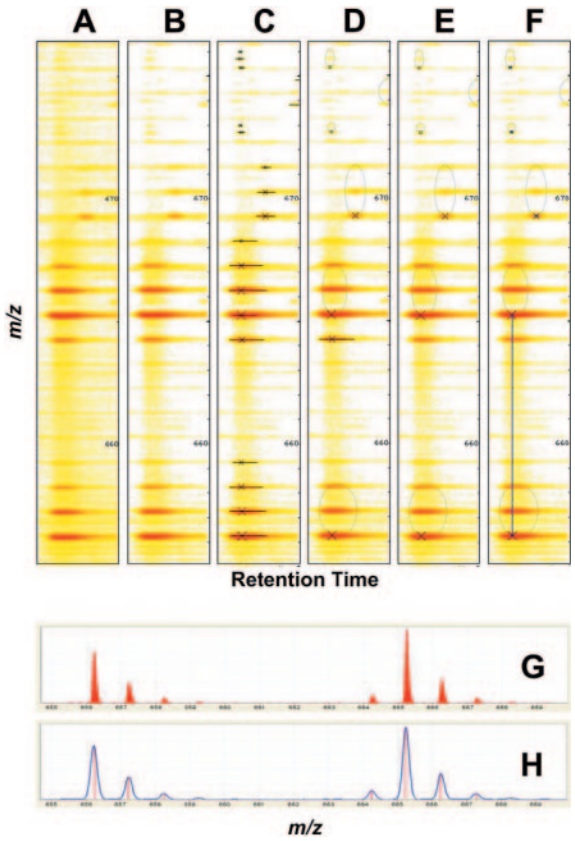

Fig. 2. Signal processing steps for peptide location and relative quantitation. (**A**) Raw spectra; (**B**) background removal; (**C**) isotope detection; (**D**) peptide cluster (isotope distribution) detection; (**E**) removal of peptides without confident charge determination; (**F**) determination of isotopically labeled pairs (vertical line joins a pair of ICAT labeled peptides); (**G**) raw data from a single scan and (**H**) peaks from G after Haar decomposition.

is built upon a number of other open-source tools including Swixml, JFreeChart, the Woodstox StAX XML parser, the Jrap mzXML parser and numerous components from the Apache Software Foundation. The msInspect executables, source code and user documentation are available at http://proteomics.fhcrc.org.

3 ALGORITHMS

Each part of msInspect contains multiple analytic steps, which are described below.

3.1 Part 1: locating eluting isotopes and peptides in individual LC-MS files

The first set of algorithms identifies the signatures of eluting isotopes in an LC-MS image (Part A) and then assembles these isotopes into peptides (Part B). The output of this step is a tab delimited peptide feature file. Each step of Part 1 may be followed using Figure 2, which shows the isotopic peaks of several peptides displayed in the msInspect graphical interface.

Part A: locating eluting isotopes in LC-MS data

Step 1: *estimate and remove local background from the LC-MS image*. The entire raw LC-MS image (Fig. 2A) is re-sampled to form an indexed image so that signal intensity at any time or *m/z* can be accessed. Background level across the image is conservatively estimated and removed using two orthogonal (time then mass), one-dimensional passes (result in Fig. 2B).

Step 2: *identify local maxima within each scan (m/z profile)*. Peaks or local maxima in each scan are identified using a wavelet additive decomposition previously described (e.g. implementation of the function 'mra' from the R statistical package) (Mallat, 1999; Percival and Walden, 2000; Randolph and Yasui, 2006). For example, the decomposed peaks in 2G are shown in 2H.

Step 3: *identify local maxima that appear to be eluting isotopes*. The peaks are smoothed over time by taking advantage of the two-dimensional nature of the image. This process identifies peaks that are sustained over multiple scans. The black lines in Figure 2C show the candidate eluting isotopes, with their maximum indicated by an 'x'.

Part B: *Assembling isotopes into peptides*. We assemble all the isotopes into groups that appear, maximize and then disappear at similar times, an indication of an eluting isotopic distribution potentially from the same peptide. The co-eluting isotopes and their observed intensities are then assembled into peptides based on their match to naturally occurring isotopes, as predicted from a simple Poisson distribution. We use a simple Poisson distribution because of its utility for modeling rare events (the rare events here being the rate of occurrence of heavy isotopes). The Poisson rate of 1/1800 was chosen based on its fit to the theoretical isotopic distributions calculated from 539 957 tryptic peptides from the human proteome sequence database (see Supplementary Figure S2). We use the Kullback–Leibler deviance (KL) (Kullback and Leibler, 1951) to compare the closeness of the observed and expected isotopic distributions. The choice of KL as a deviance score comes from many of its theoretical properties, especially its properties for measuring the discrimination information for two distributions. Intuitively it can be interpreted as a sum of penalty weights (see formula below) that will be more tolerant to deviations in lower intensity isotopes, where the relative contribution of noise is greater.

Step 1: *Assemble isotopes that appear, maximize and disappear at the same points in time and form all potential isotopic distributions*. The maximum intensity of an isotope with mass m and charge z is denoted by $I(r)$, $r = m/z$, and the $d-1$ isotopes having higher *m/z* values and within a tolerance of $(r+x)\times z$, $x = 1,\ldots,(d-1)$ are selected. By default we choose $d = 6$; if fewer than six co-eluting isotopes are identified, the remaining are assigned the value of the background intensity. An observed isotopic distribution (OID) of this candidate peptide is formed by the following:

$$\hat{P}_{rz}(x) = \frac{I(r+x/z)}{\sum_{x=0}^{d-1} I(r+x/z)}, \quad \text{for } x = 1,\ldots,(d-1).$$

Step 2: *Compare each OID to the theoretical expected isotopic distribution (EID) of a natural peptide of the same mass*. We approximate the EID using a single parameter truncated Poisson distribution given by the closed form analytic expression:

$$\mathrm{P}_m(x) = \frac{1}{K_d}\frac{(\lambda m)^x}{x!}\exp(-\lambda m), \quad x = 0,\ldots,(d-1)$$

where $\lambda = 1/1800$ and the constant K is a normalizing constant to assure the distribution sums to unity when $d < \infty$ (see Results for

discussion). Once the EID and OID are computed, we compare the deviance between the EID for $m = r \times z$ and OID using the KL (Kullback and Leibler, 1951) defined by the following:

$$\mathrm{KL} = \sum_{x} \underbrace{\hat{P}(r + x/z)}_{\text{Weight}} \underbrace{\log\left(\frac{\hat{P}(r + x/z)}{P_m(x)}\right)}_{\text{Penalty}}$$

Step 3: *Assign isotopes to a peptide based on their quality scores*. Beginning first with the peptide candidates with the best KLs, isotopes are assigned to peptides and then removed from the cluster. The process repeats until all isotopes are assigned. Unassigned isotopes are given a charge state of zero. Figure 2D shows the peptides (ellipses) located in this image, and 2E shows the charge equal to 0 peptides filtered out. The assignment of isotopes to peptides is not strictly based on KL when multiple possible assignments each yield a high quality score (low KL); assignments are biased toward peptides of lower *m*/*z* and higher charge state when multiple low scores fall within 10% of each other. This mechanism is intended to cope with the fact that isotopic distributions beginning with the second (or third) isotope in a peptide also approximate natural distributions well and may, owing to noise or natural variance, result in low KLs on their own (see Section 4.2). The same issue holds when comparing potential assignments with two different charge states: often every other isotope in a doubly-charged peptide can have an intensity distribution signature that appears similar to a natural singly charged peptide of half the mass. The bias toward smaller *m*/*z* or higher charge helps to account for the possibility of spectral noise in peak intensities resulting in a better KL for the incorrect mass or charge assignment.

The performance of the peptide location component is demonstrated by the larger mass peptide shown in Figure 2G and H. The second peak in the cluster on the right, which has a mass exactly 9 Da greater than its lighter labeled pair on the left, is the monoisotopic peak. The confounding first peak is due to incomplete incorporation of C13. Our deviance measure, KL, calculated from this confounding first peak is 1.818, but the KL from the true monoisotope, the second peak, is only 0.004. The KL computed from the third and fourth peaks are 0.062 and 0.281, respectively.

Step 4: *Quantitate each peptide and optionally combine multiple charge states*. Quantitation uses only the highest intensity peak within a peptide (see Discussion) but may be reported as either (1) the maximum intensity, (2) the intensity summed over the entire elution profile or (3) the intensities summed over the multiple charge states identified for that peptide. The option to combine multiple charge states of the same peptide (deconvolution) is available.

Following the final step, msInspect produces a peptide feature file, which lists each located peptide, its charge state(s), the time of maximum intensity, the signal intensity and several measures of quality including the KL, number of isotopes identified in the peptide and the first and last scans at which the peptide was observed.

3.2 Part 2: interrogation of individual LC-MS files

Visual curation of peptide feature files using observed isotopic profiles. The isotopic distribution for each peptide in a peptide feature file can be extracted from the raw mzXML for visual inspection of the results. Users can identify peptides that do not have the correct isotopic shape (e.g. due to misidentified charge) and may delete, correct, or annotate that feature in the peptide feature file. For a peptide located at (t, m/z) we use the signal intensities between (t, $m/z - 5$) and (t, $m/z + 5$) to construct the vector Y. Use M to denote the vector of centered mass to charges (-5, $+5$). A plot of M versus Y reconstructs the isotopic distribution. We view peptides of the same charge together in heat maps (Fig. 1D). Each sub-pane plots the identified peptide mass (horizontal axis) versus M (vertical axis) by Y (color). All correctly located peptides should have their monoisotope (first peak) centered at M = 0 with other peaks $1/z$ units apart. Users may select from the heat map those peptides that do not follow this basic pattern, and msInspect brings that peptide into focus (e.g. Fig. 1B and 1C) for closer visual inspection.

Identification of isotopic pairs. Routines are provided that locate all peptides within a single peptide feature file that (1) have the same charge state; (2) start, end and maximize at approximately the same times and (3) have a mass consistent with the mass difference between the 'heavy' and 'light' forms of the isotopic label. These isotopic pairs are identified and their intensity ratios are recorded. Figure 2F shows the location of an isotopic pair in the msInspect graphical interface.

3.3 Part 3: supporting label-free quantitation

Label-free quantitation involves combining peptide feature files from multiple LC-MS experiments into a peptide array. Much like a genomic array, rows correspond to peptides with one set of columns for every LC-MS measurement. Alignment of two or more peptide feature files involves first mapping the retention times onto a single scale and then registering peptides based on how closely their masses and mapped retention times match. To align n feature files the user first selects one peptide feature file as a reference, and msInspect creates non-linear mappings from the $n - 1$ remaining peptide feature files to this reference file. Matches are based on the closeness of peptides in the mass dimension and on this common time scale. The resulting peptide array may vary slightly depending on the choice of the reference file but not on the order of any of the other files. After the procedure has been completed, the peptide array can be normalized.

Step 1: *Estimate a non-linear mapping between the retention times of LC-MS runs*. We use iterative robust regression to estimate a transformation $f(\cdot)$ that maps the retention times t_1 from file 1 to the times t_0 from file 0, the reference file. To begin, we create matches based on mass alone from the most intense features (e.g. peptides above median intensity) and estimate a linear mapping to predict t_0 from t_1. Robust regression is used so that $f(\cdot)$ can be estimated in the presence of matching errors. We estimate the quality of the prediction using the following:

$$S_f = \sum_{(t_0, t_1)} \min\left[\left(\frac{f(t_1) - t_0}{\sigma}\right)^2, 1\right],$$

where the sum is over all pairs of times t_0, t_1 and σ reflects the scale of departure between true matches. We then iterate to estimate non-linear forms of $f(\cdot)$ by applying smoothing-spline regression methods from the previous model residuals (Huber, 1979; Hastie and Tibshirani, 1990). See Supplementary Materials for more details. This procedure is repeated for each file aligned to the reference.

Step 2: *Create matches based on closeness of mass and mapped retention times*. After Step 1, we apply the mapping to transform retention times from all files to those of the reference file. We then perform global alignment by applying divisive clustering (Kaufman and Rousseeuw, 2005) separately to both the mass and retention times of all peptides, and we select the level of clustering based on user-supplied tolerances for the cluster diameter. Peptides are aligned together, or registered, if they fall within the same cluster.

Step 3: *Optionally use a dynamic procedure to choose the optimal retention time window*. Unlike instrument mass errors, the overall tolerance used to define a match in retention time may vary from experiment to experiment. We use procedures to dynamically identify the correct tolerance for the retention time clustering. We define the quality of an alignment by the number of perfect clusters, or the number of clusters which include one and only one peptide from every experiment. msInspect may optionally iterate the retention time tolerance from small to large and select the cluster sizes to maximize the number of perfect clusters.

Step 4: *Normalize the intensities*. We implemented a normalization procedure described by Wang *et al*. (Callister *et al*., 2006; Wang *et al*., 2006b) to normalize signal intensities across multiple peptide feature files. We first extract the top order statistics from each LC-MS experiment (i.e. top intensity peptides for each run) then create a linear model using the log-intensity to predict the intensity quintiles between two runs. Order statistics are used rather than moments such as the mean and variance due to the data-dependent missing values that may occur when instrument signal intensities vary across runs. These general procedures have been shown to remove systematic biases that can be introduced due to experimental artifacts (Callister *et al*., 2006; Wang *et al*., 2006b).

4 IMPLEMENTATION AND RESULTS

4.1 Data and processing

We used three datasets to demonstrate the overall functionality and performance of the platform. Two datasets interrogate independently prepared identical aliquots of undepleted human sera using either a label-free approach or Isotope Coded Affinity Tag (ICAT) (Gygi *et al*., 1999) labeling. Identical aliquots are used to establish the ability of the platform to recover reproducible signatures from related samples, a basic component of performance that governs the ability to locate peptides that may differentiate two samples. We used msInspect to generate a peptide array from the third dataset and then used general procedures for array analysis to classify the samples.

Human serum. To evaluate label-free approaches we performed 10 independent trypsin digests of identical serum aliquots. To evaluate labeling approaches, we independently labeled three pairs of identical aliquots with heavy and light ICAT isotopes. Prepared samples were interrogated on an LCT-Premier (Waters) ESI-TOF instrument. Note that because sample preparation was performed independently, the variability among the label-free and ICAT measurements includes contributions from specimen processing. Raw data were converted to mzXML format and processed by accessing the msInspect functions via the command line using a GNU/Linux computer (3.4 GHz Intel Xeon processor, Sun 1.5.0_06 JVM with 512 GB allocated for heap). Each file was acquired over a 120 min period and averaged 6595 scans and 4.6 GB. Processing time required 25 min per file plus 38 s for alignment and normalization.

Bacteria. Two bacterial strains, *Francisella novicida* U112 (FN01, wild type) and an isogenic mutant in virulence regulator *mglA* (FN11, mglA), were processed as described in the Supplementary Material to isolate the membrane-bound protein fraction. The samples were then subjected to trypsin digestion prior to interrogation by LTQ-FT MS (Thermo Finnigan). Each strain was evaluated four times by LC-MS. Raw data were converted to mzXML format and processed by accessing the msInspect functions via the command line using a GNU/Linux computer (3.4 GHz Intel Xeon processor, Sun 1.5.0_06 JVM with 512 MB allocated for heap). Each file was acquired over an 80 min period and averaged 2336 scans and 48 MB. Processing time averaged 8 min per file plus 26 s for alignment and normalization.

4.2 Signal processing and peptide location performance

The performance of the peptide location component over a large number of peptides may be seen with the visualization/curation graphical interface. Figure 1D shows one example from the label-free serum data. The visual character of the heat map speaks to the overall validity of the signal processing method: the monoisotopic peaks line up, bands are $1/z$ apart and the thickness of the band is indicative of a parts-per-million mass tolerance.

For high-throughput use, we use KL for filtering peptides. Thresholds for accepting a good KL may vary by the number of isotopes identified and by the signal intensity. Although no systematic study of optimal filtering has been performed, we have found that removing all peptides with fewer than two isotopic peaks and $KL > 1$ is conservative, meaning that relatively few confidently located peptides (by visual inspection) will be removed. Filtering out peptides with $KL \leq 0.1$ is too strict and will remove even highly confident peptides (see Supplementary Materials).

Supplementary Table S1 summarizes the number of peptides located for each of the 10 label-free experiments. The total number of peptide candidates found before any filtering was between 6648 and 9020 (see Table S1, column 2), with a median of 7857 peptides. Filtering out peptides with $KL > 1$ reduced that number to between 5109 and 7333 (median = 6391). Subsequent combination of multiple charge states (deconvolution) further reduced that number to between 4481 and 6209, median of 5416.5, located peptides per LC-MS measurement.

4.3 Alignment and label-free quantitation

We used msInspect to merge all 10 LC-MS measurements into a peptide array for label-free profiling. To determine the specific retention time tolerance allowable when registering peptides across these runs (e.g. cluster diameter), msInspect automatically created peptide arrays for a large range of tolerance windows. For these data the optimization routine determined that a cluster diameter of 100 scans would maximize the number of unambiguous, or perfect, matches.

Figure 3 shows the resulting retention time mapping of runs 2–10 onto run 1. The common retention time scale is shown on the horizontal axis and the vertical axis shows how much each individual run was adjusted. The black line represents the retention time mapping from the first run to itself, and the remaining curves show the mapping for all other LC-MS runs to this first run. The

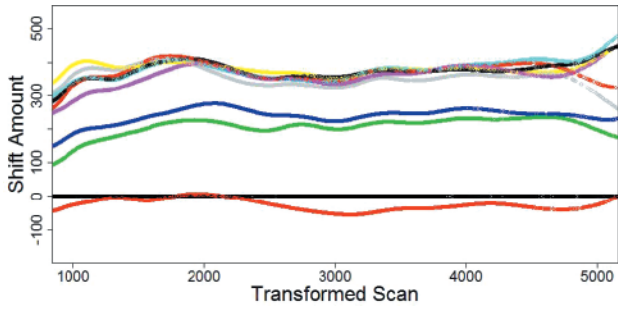

Fig. 3. Shift of peptides during the retention time mapping process. Horizontal axis is the scan number of each peptide after mapping. Vertical axis is the amount of 'shift' applied to transform the scan number of each peptide during retention time mapping. Each run is indicated by a different color.

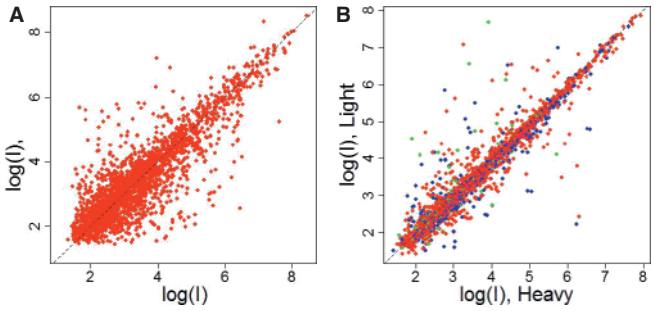

Fig. 4. Performance of quantitation using (**A**) label-free and isotopic (**B**) labeling approaches. Point color indicates the number of isotopic labels detected, red for one label, blue for two, and green for three.

distance between the curves and the line can be represented by a linear shift and a remaining non-linear component. The magnitude of the reduction should be measured in relation to the final cluster size, or tolerance, chosen. For the sera data, where a final retention time deviation of 100 scans was used, the linear shift reduced the retention time deviation by a median of 340 scans, or 340%, and the non-linear component further reduced the required deviation by an additional 36 scans, or 36% of the resulting tolerance.

To evaluate the ability to recover reproducible signatures when using label-free approaches, we examined the signal intensity correlation across pairs of experiments. Across all pairs, the Pearson correlations fall between 0.816 and 0.956 (mean = 0.905). Figure 4A summarizes this correlation graphically by plotting the log intensities for one of the 45 possible pairwise comparisons. Fewer than 20% of all intensity ratios exceed a 2-fold change. Note that because each interrogation was performed on serum samples that were independently processed (e.g. protein digestion) these reproducibility measurements contain multiple sources of variation, including biochemical, instrumentation and data analysis. The reproducibility shown here results from completely automated analysis procedures and without strict filtering. Overall reproducibility may improve by using more stringent filtering criteria (e.g. requiring three identified isotopes), by using visual curation to eliminate errors in feature detection, or perhaps by addressing the reproducibility of specimen handling and biochemical procedures. The unfiltered data for these experiments are provided in Supplementary Material.

4.4 Quantitation using stable isotopes

To establish reproducibility using isotopic labeling approaches, we analyzed ICAT-labeled serum using msInspect and filtered to remove peptides with KL > 1 and fewer than two isotopes. Peptides were paired by assigning cleavable ICAT label weights (light = 227.126, heavy = 236.156, with up to three labels) and assuming a mass tolerance of 50 p.p.m. The numbers of peptides (and labeled pairs) found in the six replicates were 8504 (2035), 7929 (1930), 5606 (1331), 4949 (1209), 8837 (2093) and 7926 (1911) from run one to six, respectively. Figure 4B shows the intensities of the 'light' peptides plotted against the intensities of the corresponding 'heavy' peptides, with color indicating the number of labeling tags detected for each pair; the remaining five plots for the ICAT-labeled serum data are shown in the Supplementary Material Figure S3. The correlations of the six runs, a measure of overall reproducibility, were 0.930, 0.954, 0.952, 0.978, 0.937 and 0.916 from run one to six, respectively. Over 90% of all isotopic pairs found had ratios within 1.5-fold change. Unfiltered data are provided in the Supplementary Material.

4.5 Application of msInspect to molecular profiling

The examples above demonstrate the performance of individual msInspect components. Here we use the LC-MS measurements of two strains of bacteria to demonstrate the ability to correctly classify two different biological samples using msInspect, including one sample that was affected by an unplanned experimental artifact.

Following peptide location and filtering (deconvolution was not applied), we found 1554, 1694, 1644 and 1628 peptides in the four mutant samples (MUT) and 1437, 1372, 1313, 1867 peptides in the four wild-type runs (WT). We performed alignment and normalization using the procedures described above. We first evaluated the similarity of the LC-MS measurements. The signal intensities were more highly correlated within strains than between. Pearson correlations within the WT and within the MUT groups averaged 0.933 and 0.951, respectively, but correlations between strains averaged only 0.851.

We next subjected the peptide array to unsupervised clustering following some modest filtering which eliminated all peptides that were found in only one of the eight LC-MS measurements. A total of 2024 peptides remained and were clustered using the Hierarchical Clustering command *hclust* from the R statistical package. The results are shown in Figure 5. The dendrogram shows a strong association was found within each of the WT and MUT strains as well as a large degree of difference between groups.

Moreover, inspection of the dendrogram also shows that one measurement, the fourth WT, was less similar to the other WT samples even though it clustered strongly with them overall. We used the retention time mapping plots (Supplementary Figure S4) to investigate this observation. The overall experimental design interrogated samples WT1 through WT4 then MUT1 through MUT4. Figure S4 reveals that just prior to the interrogation of WT4 the gradient lengthened and the overall signal intensity dropped, an unplanned and previously unknown occurrence. The effect of this experimental artifact was to make the WT4 measurement artificially more similar in retention time and signal intensity to the MUT groups than to the other WT group members. Despite this artifact, msInspect graphical features detected this anomaly, retention time mapping and registration adjusted for the gradient length,

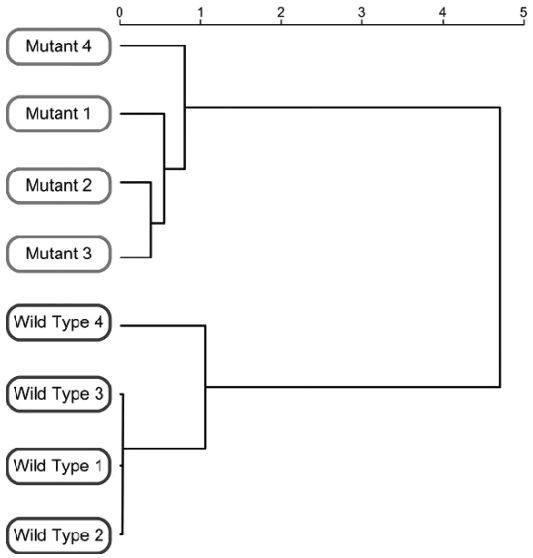

Fig. 5. Dendrogram resulting from the unsupervised clustering of the bacteria data.

and normalization adjusted for the change in signal intensities enough to allow this sample to be correctly classified with the other members of its biologically similar group.

DISCUSSION

For the examples presented here, we demonstrate the ability of msInspect to identify reproducible signatures from complex samples using both label-free and isotopic labeling approaches. Although it is not possible to comprehensively evaluate each component, we gave several examples that demonstrate the overall ability of msInspect to locate and quantitate peptides and compare complex mixtures. These examples included: (1) the 'heat map' visualization, which demonstrates that nearly all peptides found contain the isotopic pattern expected given their charge state and mass; (2) the high correlation of intensities from the labeled serum samples, which demonstrates the general ability to quantitate those peptides; (3) the correlation of intensities from the label-free approach, which demonstrates the ability to map retention times and register peptides across multiple LC-MS measurements and (4) the molecular classification of two strains of bacteria, which demonstrates the overall ability to put all components together to address biologically relevant questions.

As we developed the algorithms for msInspect, we made several choices based on knowledge of basic quantitative principles and routine visual inspection of data. For example, of the many quality measures we could use to improve peptide location, we chose to compare resulting distributions to a simple Poisson approximation to identify natural isotopic shapes. We found the Poisson approximation performed as well as the more computationally-intensive approaches described by Gay *et al*. (1999) but was more accurate when used outside the range they considered. Also, when assigning a quantity to a located peptide we chose to use only the maximum intensity of its isotopes rather than summing intensities over all of its identified isotopes. We made this choice because the most intense peaks are typically those measured with greatest precision, and a simple summation that includes the less-intense, less-precisely measured isotopes could result in a reduction in overall precision. A related argument holds when considering the option of summing multiple charge states. Simply summing different components, each measured with different precision, could result in reduced reliability overall. Thus, although we provide a mechanism to combine multiple charge states by summing their intensities, we do not recommend doing so until better procedures for combining multiple charge states are developed.

It is unlikely that any single implementation of the entire pipeline outlined by Listgarten and Emili (2005) will contain the best approach at every step, and continued research will be needed for each of the individual components. The best step forward may come from integrating and comparing the different open-source algorithm implementations (Katajamaa, *et al*., 2006; Katajamaa and Oresic, 2005). The msInspect program has been designed to allow replacement of many individual components with alternate, competing methods, and, like other approaches with different algorithms, the source code has been made available in order to foster the synthesis of the best components into one or more platforms.

ACKNOWLEDGEMENTS

This work was funded by National Cancer Institute subcontract 23XS144A. Funding to pay the Open Access publication charges was provided by the National Cancer Institute subcontract 23XS144A.

Conflict of Interest: none declared.

REFERENCES

Adam,B.L. *et al*. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*., **62**, 3609–3614.

Callister,S.J. *et al*. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome. Res*., **5**, 277–286.

Domon,B. and Aebersold,R. (2006) Mass spectrometry and protein analysis. *Science*, **312**, 212–217.

Etzioni,R. *et al*. (2003) The case for early detection. *Nat. Rev. Cancer*, **3**, 243–252.

Gay,S. *et al*. (1999) Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis*, **20**, 3527–3534.

Gygi,S.P. *et al*. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol*., **17**, 994–999.

Hastie,T.J. and Tibshirani,R.J. (1990) *Generalized Additive Models*. Chapman and Hall, New York.

Huber,P. (1979) Robust Smoothing. In Launer,R.L. and Wikcinson,G.N. (eds), *Robustness in Statistics*. Academic Press, New York.

Katajamaa,M. and Oresic,M. (2005) Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, **6**, 179.

Katajamaa,M. *et al*. (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, **22**, 634–636.

Kaufman,L. and Rousseeuw,P.J. (2005) *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley, Hoboken, N.J.

Kullback,S. and Leibler,R. (1951) On information and sufficiency. *Ann. Math. Stat*., **22**, 79–86.

Li,X.J. *et al*. (2005) A software suite for the generation and comparison of Peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell Proteomics*, **4**, 1328–1340.

Listgarten,J. and Emili,A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell Proteomics*, **4**, 419–434.

Mallat,S.G. (1999) *A Wavelet Tour of Signal Processing*. Academic Press, San Diego.
Morris,J.S. *et al*. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, **21**, 1764–1775.
Pedrioli,P.G. *et al*. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol*., **22**, 1459–1466.
Percival,D.B. and Walden,A.T. (2000) *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge, NY.
Petricoin,E.F. *et al*. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**, 572–577.
Randolph,T.W. and Yasui,Y. (2006) Multiscale processing of mass spectrometry data. *Biometrics*, **62**, 589–597.
Smith,R.D. *et al*. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics*, **2**, 513–523.
Vlahou,A. *et al*. (2001) Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. *Am. J. Pathol*., **158**, 1491–1502.
Wang,P. *et al*. (2006a) A statistical method for chromatographic alignment of LC-MS data. *Biostatistics*, (Under Review).
Wang,P. *et al*. (2006b) Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Proc. Pac. Symp. Biocomput*., **11**, 315–326.
Yasui,Y. *et al*. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, **4**, 449–463.