Signal Maps for Mass Spectrometry-based Comparative Proteomics*

Amol Prakash‡§¶, Parag Mallick‖, Jeffrey Whiteaker¶, Heidi Zhang¶, Amanda Paulovich¶, Mark Flory**‡‡, Hookeun Lee§§, Ruedi Aebersold**§§, and Benno Schwikowski‡§**¶¶**

Mass spectrometry-based proteomic experiments, in combination with liquid chromatography-based separation, can be used to compare complex biological samples across multiple conditions. These comparisons are usually performed on the level of protein lists generated from individual experiments. Unfortunately given the current technologies, these lists typically cover only a small fraction of the total protein content, making global comparisons extremely limited. Recently approaches have been suggested that are built on the comparison of computationally built feature lists instead of protein identifications. Although these approaches promise to capture a bigger spectrum of the proteins present in a complex mixture, their success is strongly dependent on the correctness of the identified features and the aligned retention times of these features across multiple experiments. In this experimental-computational study, we went one step further and performed the comparisons directly on the signal level. First *signal maps* were constructed that associate the experimental signals across multiple experiments. Then a feature detection algorithm used this integrated information to identify those features that are discriminating or common across multiple experiments. At the core of our approach is a score function that faithfully recognizes mass spectra from similar peptide mixtures and an algorithm that produces an optimal alignment (time warping) of the liquid chromatography experiments on the basis of raw MS signal, making minimal assumptions on the underlying data. We provide experimental evidence that suggests uniqueness and correctness of the resulting signal maps even on low accuracy mass spectrometers. These maps can be used for a variety of proteomic analyses. Here we illustrate the use of signal maps for the discovery of diagnostic biomarkers. An implementation of our algorithm is available on our Web server.
Molecular & Cellular Proteomics 5:423–432, 2006.

The ability to probe a complex biological sample globally at the protein level is of key importance for the advancement of systems biology approaches (1), which are built on unbiased, global measurements of cellular processes. Used across multiple samples, they can be used to identify common similarities and differences. MS-based proteomic analysis is one of the most promising technologies applied toward these goals (2–5). This approach can broadly be divided into two types: MS-based fingerprinting and MS/MS-based sequencing.

Although mass spectrometers are exquisitely sophisticated instruments, there is evidence that they suffer from problems of sensitivity, reproducibility, and undersampling. The number of human genes, for example, is estimated to be well above 20,000, but protein products for only a fraction of those have been detected so far (6). The main reason is the limited capability to detect very low abundance analytes, thus highlighting the low sensitivity of these instruments, thus creating problems for the fingerprinting-based approaches. Furthermore the sample throughput of tandem MS-based methods presents another limitation of this technology. Peptides are typically selected for CID based on the intensity of the MS signal they generate. Because only some of the larger peaks are typically subjected to CID, this undersampling issue worsens the ability to profile a sample completely and confidently. This problem is further augmented due to the low sensitivity of the search algorithms used to sequence MS/MS spectra. Both of the above problems are approached (with limited success) by way of repeated experiments where the net throughput is better than that of the individual experiments. This further demonstrates the problem of reproducibility, *i.e.* the intensity of peptide peaks varies across experiments, and thus widely varying subsets of peptides get identified and quantified in each. All these issues create a fundamental problem for large scale MS-based proteomic studies, particularly those that are comparative in nature (for example, identifying biomarkers in human serum).

In the last few years, some computational approaches have been suggested to tackle these issues, for example accurate mass tags (7) and PepMiner (8). These approaches are based

From the ‡Department of Computer Science, University of Washington, Seattle, Washington 98195, **Institute for Systems Biology, Seattle, Washington 98103, ¶Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, §§Institute for Molecular Systems Biology, Eidgenössische Technische Hochschule and Faculty of Natural Sciences, University of Zurich, Zurich, Switzerland, ‡‡Department of Molecular Biology and Biochemistry, Wesleyan University, Middletown, Connecticut 06459, §Systems Biology Group, Institut Pasteur, 25-28 Rue du Dr. Roux, 75015 Paris, France, and ‖Cedars-Sinai Medical Center, Los Angeles, California 90048

Received, May 11, 2005, and in revised form, October 31, 2005
Published, MCP Papers in Press, November 3, 2005, DOI 10.1074/mcp.M500133-MCP200

TABLE I
Formal definitions of the various terms and notions used
The symbol convention used in this table is followed in the entire text.

Term used	Formal definition
Mass spectrum M	A set $\{p_1, p_2, \ldots\}$ of peaks.
Peak p	Each peak has an associated mass-to-charge ratio mz(p) and intensity(p).
Run R	A series $(M_1, M_2, \ldots)$ of mass spectra. A run corresponds to the output of an RPLC experiment.
Run T0SCX23A	A run in a two-dimensional LC/LC MS experiment. The notation denotes the series A run of SCX fraction 23 of the sample collected at time point 0.
$p \leq q$	We write $p \leq q$ for two peaks $p \epsilon M_i$ *and* $q \epsilon M_j$ *if* $i \leq j$, *i.e.* p and q occur in the same run, and p occurs in the same or an earlier spectrum than q.
Signal map f	We call f *order-preserving* if, for any pair (p,q) of peaks with $p \leq q$, there exists a peak $u \epsilon f(p)$, such that $u \leq v$ for all peaks $v \epsilon f(q)$.
Alignment α	Given two 1D runs R with spectra $(M_1, M_2, \ldots)$ and S with spectra $(N_1, N_2, \ldots)$, a function α that maps each spectrum in R to a set of spectra in S. α has the property that, for any two spectra M_i *and* M_j *with* $i \leq j$, there exists a spectrum $N_k \epsilon \alpha(M_i)$, *such that* $k \leq l$ for all spectra $N_l \epsilon \alpha(M_j)$.

on building MS fingerprints of the peptide signals. Commonly if multiple experiments on related samples are to be compared, they are first interpreted individually, resulting in lists of identified peptides or proteins, one per experiment, that are then compared. The success of these approaches is limited due to the low throughput of the tandem MS experiments and the semistochastic subset of proteins identified in each experiment. Recently several research groups (9–11) have built tools to computationally identify and list the features (putative peptides) in a single MS experiment. These lists are then compared across multiple experiments by matching the *m/z* value and the retention time of the individual features. Peptide elution is not a well understood process, and a lot of unexplained behavior is observed even in the elution profiles of the sequenced peptides, making these feature lists ambiguous (evident by different feature lists generated by different tools on the same sample). Also the retention time for a peptide is notoriously hard to reproduce (12), a notion that was confirmed by the experimental data we examined here. Reasons for this are variations in the speed with which the peptides travel through the LC column, variations in time elapsing between successive fractions, inconsistencies in the solvent gradient generation, etc. Thus computational alignment of the LC column across multiple experiments is a hard problem, but the above methods do not give evidence for the correctness of the alignments that they construct. Furthermore the above methods make additional assumptions, *e.g.* that retention time can be interpolated linearly between blocks in different experiments (10, 13), that features can be reliably recognized from single experiments (10), that the total ion count profile suffices for the characterization of the alignment in time (9, 14), and that retention time varies by about 5 min or less between experiments (11).

In this study, we suggest a novel approach that is significantly different from all the above. First a *signal map* that maps the signals from any given peptide in one experiment to the signals acquired from the same peptide in the other experiment was created. Once constructed, signal maps can then be exploited for a variety of purposes, for example, for detecting features common across large numbers of experiments. The intuition behind this approach is that feature extraction on the basis of multiple experiments will be more sensitive and specific than first identifying features and then comparing across multiple experiments. Our approach first constructs a signal map on the raw MS data and performs all other data processing (*e.g.* feature detection) later. Similar ideas are successfully used in comparative genomics where *genome maps* are first created using whole-genome alignments and then used for better understanding of the content and function of various genomic regions (15).

Table I presents an overview of the formal notions we used. Let M and N be two MS runs where M_1, M_2, . . . and N_1, N_2, . . . are the corresponding mass spectra. Our technical approach (described in detail under "Materials and Methods") constructs the signal map by constructing an alignment that consists of a set of pairs (i,j) such that the occurrence of shared peaks between the pairs M_i and N_j is globally maximized. It is a stronger version of *dynamic time warping*, a highly successful approach in natural speech processing and other fields (16). To account for varying column speeds, we allow for multiple mass spectra in one run to correspond to a single mass spectrum in the other run. Our approach assumes that *elution order* (the relative order of peptides) is conserved. This assumption is likely to hold because peptides are separated by the same physical property in both runs. All methods described earlier (10, 11, 13, 14) make this assumption without presenting direct experimental evidence for it. In this study, other than describing the usefulness of signal maps, we also present (*a*) a method that does not make any additional assumptions beyond elution order conservation, (*b*) direct experimental evidence that this assumption holds, and (*c*) direct experimental evidence for the correctness of the alignments resulting from our method even on low-to-moderate accuracy instruments.

We also present evidence that the approach of signal maps is highly sensitive at identifying even low intensity signals and thus can potentially be used for increasing throughput. A

detailed performance comparison of this approach with the other techniques requires a perfect experimental setup (every peak either has a peptide identification or is classified as noise), which is infeasible given the current technologies. Other potential uses for signal maps include exploring the multiple experiment neighborhood of a peak for its presence or absence to enhance feature recognition techniques and evaluation of reproducibility of experimental setups, *e.g.* clinical trials of drugs, column optimization, etc.

MATERIALS AND METHODS

An implementation of our algorithm is available on our web site at www.systemsbiology.fr/chams.

Score Function—Here we consider two runs $R = (M_1, M_2, \ldots)$ and $S = (N_1, N_2, \ldots)$. Generally speaking, our aim was to construct an alignment that places similar spectra M_i and N_j close to each other. Our procedure for finding such an alignment is based on a pairwise score function. Note that a spectrum is a list of peaks, so the spectrum M_i can be written as $\{p_1, p_2, \ldots\}$ where mz(p_i) and intensity(p_i) are the *m/z* ratio and intensity value for the peak p_i, respectively (see Table I).

This score function rewards corresponding peaks within a window of 2ϵ where ϵ is the accuracy of the mass spectrometer used. Our initial measure of agreement between the two spectra $M_i = \{p_1, p_2, \ldots\}$ and $N_j = \{q_1, q_2, \ldots\}$ is: $M_i \times N_j = \Sigma_{(p,q)}$ intensity(p) $\times$ intensity(q) where $|\text{mz}(p) - \text{mz}(q) \leq 2\epsilon|$. Normalizing this expression to make it robust against global linear fluctuations in peak intensity, we arrive at the following preliminary score function for two mass spectra M_i and N_j: $(M_i \times N_j)/\sqrt{((M_i \times M_i)(N_j \times N_j))}$, which is the same measure as the one used by Stein and Scott (17).

However, we found that the above score function does not appear to very well distinguish "close" from "distant" spectra. To address this, we computed an additional term $E[M_i \times N_j]$, which denotes the expected value of $M_i \times N_j$ under the random placement of all peaks within the given mass-to-charge range.

$$s(i,j) := \frac{M_i \times N_j - E[M_i \times N_j]}{\sqrt{(M_i \times M_i)(N_j \times N_j)}} \quad \text{(Eq. 1)}$$

If we consider each spectrum as being generated from two components: a noise distribution and a signal distribution, $E[M_i \times N_j]$ approximates $s(i,j)$ for M_i and N_j generated only from the noise distribution. Thus subtracting this term makes only the signal component contribute to the score. The expected value of $s(i,j)$ is thus zero for completely uncorrelated mass spectra M_i and N_j. Fortunately $E[M_i \times N_j]$ is straight forward to compute. Note that the probability that two peaks p and q with randomized intensity values are within two ϵ of each other is constant. Therefore, a constant c proportional to ϵ exists, such that

$$E[M_i \times N_j] = \sum_{(p,q)} c \times \text{intensity}(p) \times \text{intensity}(q)$$

$$= c \times \left(\sum_p \text{intensity}(p)\right) \times \left(\sum_q \text{intensity}(q)\right) \quad \text{(Eq. 2)}$$

Thus, after precomputing (Σ_p intensity(p)) for M_i and (Σ_q intensity(q)) for N_j, $E[M_i \times N_j]$ can be computed for each given pair (M_i,N_j) with only three multiplications. To further reduce the effect of unrelated spectra, any score below 0.2 was reduced to zero.

Alignment Algorithm—Based on the above score function $s(i,j)$, we aimed to relate peaks of a run $R = (M_1, M_2, \ldots)$ with peaks of another run $S = (N_1, N_2, \ldots)$ through a signal map f by choosing the alignment α such that similar spectra appear close to each other in the sequence (i,j) in α. In empirical tests, we determined that it is beneficial to introduce some robustness against bad spectra by scoring not only pairs of spectra immediately adjacent in α but also *close* ones as follows.

For each spectrum M_i, we also scored its similarity with those two spectra N_{j-1} and N_{j-2} that appear immediately before N_j in α and those two spectra N_{j+1} and N_{j+2} that appear immediately after N_j in α (omitting the border cases). N_j can also be scored against M_i in a similar way. Adding all these similarity scores gives us a much stronger estimate for the local similarity between M_i and N_j. A best alignment according to this score function can be computed using a straightforward modification of the Needleman-Wunsch global alignment algorithm (18) as described below.

Let $sc_\alpha(k,l)$ be the score of the best alignment of runs (M_1, M_2, $\ldots M_k$) and (N_1, N_2, $\ldots N_l$). Then using dynamic programming we can compute $sc_\alpha(i,j)$ by the following recursion.

$$sc_\alpha(i,j) = \max[sc_\alpha(i,j-1) + s(i-2,j) + s(i+1,j) + s(i,j) + s(i+1,j) + s(i+2,j),$$
$$sc_\alpha(i-1,j) + s(i,j-2) + s(i,j-1) + s(i,j) + s(i,j+1) + s(i,j+2)] \quad \text{(Eq. 3)}$$

Formulating the problem in the above manner, the dynamic programming computes the score $sc_\alpha(k, l)$ of the global alignment with the maximum spectra similarity, giving us a global measure of similarity between the two runs. As in sequence alignment, an optimal alignment can be extracted by backtracking the optimal values in the abovere cursion.

We validated the above approach by examining a few pair wise alignments. The data set we used represents five duplicate two-dimensional LC-MS/MS measurements of tryptic digests of a whole-cell lysate of the yeast *Saccharomyces cerevisiae.* The five samples were collected from cells synchronized in the G_1 phase of the cell cycle and at four time points following release (30, 60, 90, and 120 min). The protein sample was digested into peptides using the enzyme trypsin and then separated by charge into 35 fractions using SCX chromatography. For each time point, each fraction was split in half, resulting in a "Series A" and a "Series B," which together have a total of 68 fractions (at time point 0, data for two Series B fractions were not acquired). Each sample, in turn, was separated according to hydrophobicity by RPLC and later analyzed by ESI-MS (ThermoFinnigan LCQ Deca XP mass spectrometer). Complete detail about this data set is presented in Flory *et al.*[1] We use the notation T0SCX23A to refer to the RPLC experiment done on Series A of SCX fraction 23 of sample acquired at time point 0 min. Fig. 1 shows three alignments for this purpose. The alignment of run T0SCX23A with itself is a near perfect diagonal alignment as one would expect. The other two plots show the alignment of run T0SCX23A with runs T0SCX23B (repeat experiment) and T0SCX24A (the strong cation exchange (SCX)[2] fraction following T0SCX23A). None of these alignments is a perfect diagonal, but the middle region of the run where (from our more detailed examination) almost all of the signals lie is quite close to diagonal (similar evidence in Figs. 5 and 8). As expected, the technical replicate experiment T0SCX23B had an alignment closer to diagonal than the adjacent SCX fraction T0SCX24A, which could be consid-

[1] M. Flory, H. Lee, R. Bonneau, P. Mallick, K. Serikawa, D. R. Morris, and R. Aebersold, submitted for publication.

[2] The abbreviations used are: SCX, strong cation exchange; RP, reverse phase; ID, identification; AnSi, analysis of similarities; AnDi, analysis of differences.

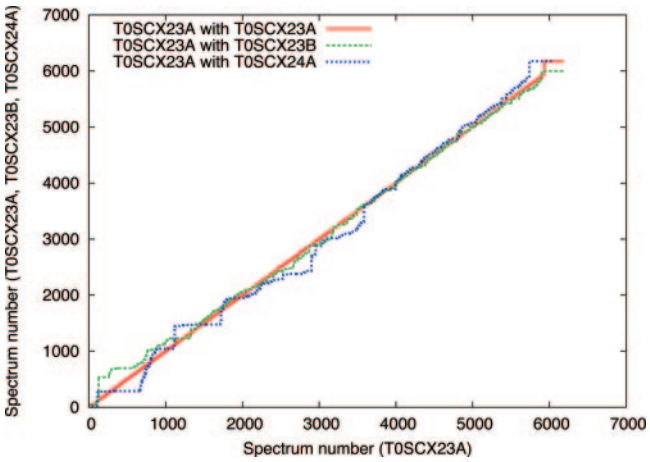

FIG. 1. **The alignment of run T0SCX23A with runs T0SCX23A, T0SCX23B, and T0SCX24A for yeast whole-cell lysate experiment.**

ered as an approximate "biological replicate." It is important to note that the similar, but over a large interval, constant and distinct, slopes of the alignments are in no way favored by the algorithm itself. This observation suggests that the alignment reveals a true difference in the relative speed with which the peptides elute in the seruns. This phenomenon is known among mass spectrometrists as a difference in flow rates of the two reverse phase (RP) LC columns, and flow rate is one parameter that is known to be hard to keep constant between runs. This empirical observation confirms the ability of our optimal alignment to reconstruct the correspondence between two runs of mass spectra.

To study whether these alignments are at all necessary or a diagonal alignment might be of a similar quality, we compared the scores of the alignments generated by our algorithm to the perfectly diagonal alignments. The score of the optimal alignment of T0SCX23A with T0SCX23B was 2638, and the score for the perfectly diagonal alignment was 159. The corresponding scores for the alignment of T0SCX23A with T0SCX24A were 3300 and 472, respectively. These dramatic differences suggest that the assumption of simple perfect diagonal alignments cannot adequately group similar spectra of different runs together.

Analysis of Similarities and Analysis of Differences—Both these analyses required us to analyze multiple (more than two) runs together. Based on our method for constructing pairwise signal maps by alignment, we explored two strategies.

Global Alignment Using a Set of Globally Best Pairwise Alignments—The pairwise alignment procedure developed above allowed us to map the signals between any pair of 68 runs (of time point 0) in our input data (yeast experiment) and thus match signals between successive runs. To extend the pairwise map to a global map, we needed to thread together more than two runs. Although this might be done in the linear order in which the runs were acquired, we tried the following approach to guard against potentially bad runs. We decided to maximize the quality of the set of the pairwise alignments that we can choose to thread together to form a global alignment.

Given $n = 68$ runs, we would always need $n - 1 = 67$ pairwise alignments to connect all runs into a single spanning graph. For this, we decided to compute a minimum spanning tree (19) of the complete graph that contains all pairwise alignments. Fig. 2 shows the minimum spanning tree.

We consider it a remarkable testimony to the quality of the data, our score function, and alignment algorithms that, without providing any prior information on the historical order of the runs, the edges of the resulting spanning tree have such a high degree of agreement with the historical order. This outcome strengthens our confidence in the alignment scoring method.

Progressive Multiple Alignment—The other approach we proposed is based on a progressive alignment strategy. For this, we computed all pairwise alignments. Then the strongest pair (having the highest alignment score) was merged to form a *consensus* run (using the analysis of similarities (AnSi) approach described below). This procedure was then repeated to identify the next pair of runs to merge. The process was repeated until we were left with a single consensus run.

AnSi—To analyze the similarities between two runs, we merged the two into a single run with the idea that the similarities would be strengthened. The merge operation follows the signal map between the two runs. For each pair (M_i,N_j) of MS spectra in the alignment α, we generated a new spectrum by overlaying the two spectra. Peaks close in *m/z* value were merged into one by adding the intensity and averaging the *m/z* value (weighted by their intensity). In this way we could create a run of merged spectra, which could then be treated as a consensus run.

Analysis of Differences (AnDi)—To identify differences between two runs, the merging happens in a slightly different way. Instead of merging the two spectra (M_i,N_j) together, for each peak of M_i we searched for the highest corresponding peak in a small window ($\pm$10) of spectra around N_j. The window allows for small errors in the alignment and a not well understood elution profile. This largest peak was then used to normalize the corresponding peak in M_i. This way we got the difference spectrum M_i. A run of such spectra $(M_1, M_2, \ldots)$ gives the consensus difference run.

Feature Recognition Method—To detect features in real and virtual runs, we used the following simple feature recognition algorithm. We called a peak a *feature* if the following three conditions are satisfied.

- The peak has a certain intensity threshold (5-fold is used for our study).
- Most (80%) of the corresponding peaks in the nearby ($\pm$5) spectra have high intensity.
- There is another peak within the isotopic range of this peak.

Peptide ID Transfer—Two CID spectra were compared using the following three factors, and if they were significantly similar, the identification was transferred from one to the other.

- Having similar precursor masses ($\pm$0.5Da).
- Having a high similarity score (defined earlier as $s(i,j)$).
- Having a high similarity score even after we remove the top five peaks to increase the effect of low intensity peaks.

If the two CID spectra showed high similarity on the above three factors, the identification was transferred from the source to the target. To do an interexperiment peptide transfer, all CID spectra in one experiment were compared with the ones from the second experiment using the above procedure.

In addition to the above criteria, we could also add proximity in signal map, *i.e.* the precursor peaks p and q in the two runs (which correspond to the same peptide) are such that $f(p)$ and q are close to each other in spectra indices. As described earlier, we first computed all pairwise alignments. Then using this signal map consisting of all *strong* pairwise alignments (alignments with average spectra similarity greater than 0.2), we could decide whether the two CID spectra are close. As capturing peptides for CID happens on an irregular basis, we used a relaxed threshold ($\pm$100 spectra) to decide whether the two CIDs refer to the same peptide.

RESULTS

Our Score Function Recognizes Similar Peptide Mixtures—At the basis of our method is the score function $s(i,j)$

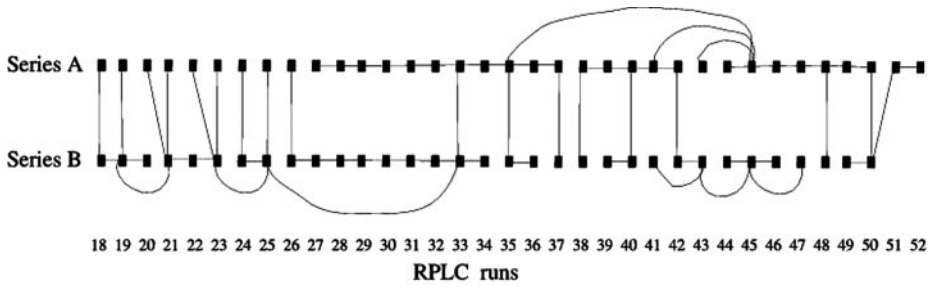

FIG. 2. **Minimum spanning tree for the 68 runs in time point 0 for yeast whole-cell lysate experiment.** The various connections show the high scoring alignments.

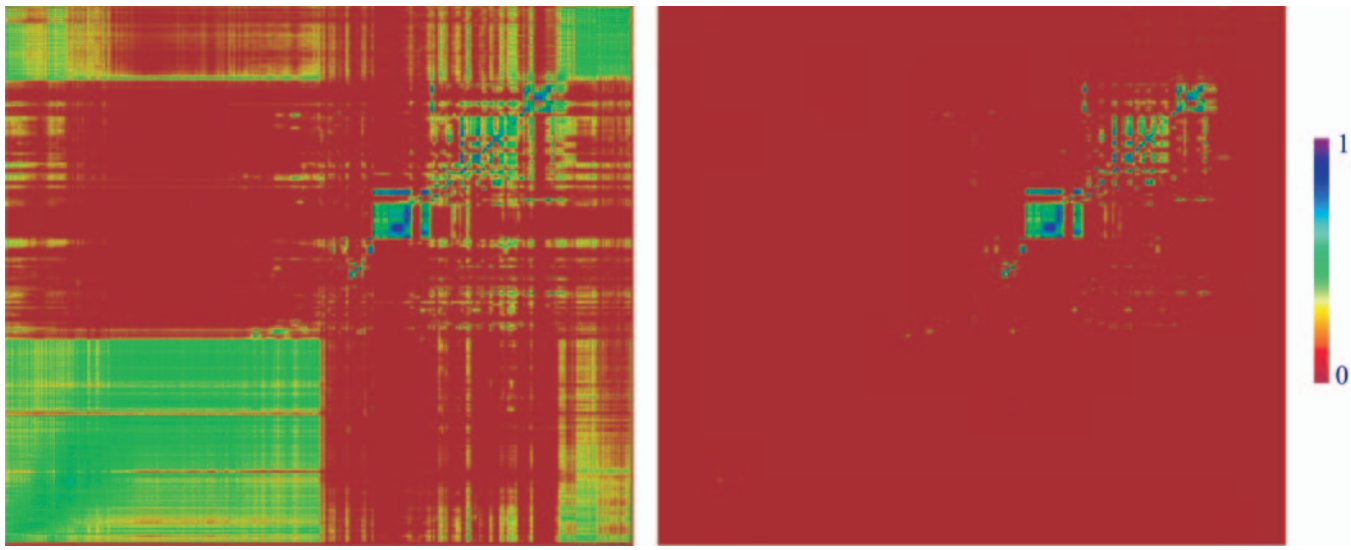

FIG. 3. **The value of the Stein and Scott (17) score *s(i,j)* (*panel 1*) and our score *s(i,j)* for various spectrum pairs (*i,j*) (*panel 2*).** Indices *i* and *j* label the *horizontal* and *vertical axes*, respectively, where M_i and N_j are mass spectra from the two runs T0SCX23A and T0SCX24A, respectively (representing adjacent SCX fractions in the yeast whole-cell lysate experiment). The color code is shown on the *right*. As expected, our score function gives lower scores to unrelated spectra and thus is better able to classify similar spectra from dissimilar ones.

that quantifies the amount of shared peaks in any pair (M_i,N_j) of mass spectra. As described in "Materials and Methods" the score is based on the number of peaks close in *m/z* value (considered close enough to come from the same peptide) and their intensities.

To evaluate how well score functions work in this context, we considered pairs of successive SCX fractions of time point 0 from the yeast cell cycle data. These can be expected to have some overlap in peptide content, and high scores should thus be generated for those spectrum pairs (M_i,N_j) with similar *i* and *j*. When scores *s*(*i*,*j*) are represented in rectangular array with coordinates (*i*,*j*), high scores are thus expected around a path from low (*i*,*j*) pairs to high (*i*,*j*) pairs (in the case of a perfectly linear relationship between elution times, around a diagonal).

Fig. 3 represents such arrays for the runs T0SCX23A and T0SCX24A on the *left-hand side* for the Stein and Scott (17) score function that performed best in a previous comparison of score functions and on the *right-hand side* for our score function. In both cases, high scores did occur around portions of the diagonal, but the Stein and Scott (17) score function yielded many more high scores for two more classes of *unrelated* spectra for which scores should be low: (*a*) those that are off-diagonal and not generated from similar spectra and (*b*) those pairs generated from the beginning and end of the two experiments. Upon closer inspection, we established that the beginning and the end of both runs contain long subseries of "empty" spectra that contain no significant signal. Similar results were obtained when other pairs of successive SCX fractions were analyzed.

As another assessment of the quality of our score function, we tried to quantify how well our score function can distinguish pairs of spectra from similar peptide mixtures. As before, spectra from similar peptide mixtures are the pairs that high scoring alignments should bring together; hence these pairs need to receive a significantly higher score than pairs of spectra that are generated from unrelated spectra. To come up with approximate positive and negative test sets for this discrimination requirement, we used the run T0SCX23A and called two mass spectra M_i and M_j *related* whenever $|i - j| < 10$ and unrelated otherwise. As peptides have an elution profile spanning multiple successive spectra, nearby spectra in a run are usually generated from similar peptide mixtures. We then tested how well our score function was able to

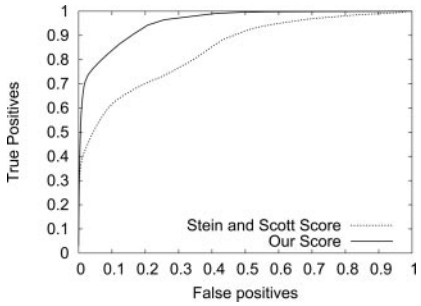

FIG. 4. **Receiver operating characteristic curve comparing our score function with the score function from Stein and Scott (17) on the run T0SCX23A.**

classify pairs of spectra into these two classes only based on the peak data itself. Again we compared with the Stein and Scott (17) score. The resulting receiver operating characteristic curves for both score functions on the same data set are shown in Fig. 4. It shows that, at all thresholds, our score function identified a higher number of true positives and fewer false positives when compared with the Stein and Scott (17) score. This indicates that our score function represents a significant improvement in specificity and sensitivity for the identification of similar peptide mixtures. Similar results were obtained when we varied the threshold used for classifying $|i - j|$ as related or used other runs instead of T0SCX23A.

Optimal Signal Maps Are Well Defined, and They Expose Local Warps—Having established that our score function for pairs of mass spectra recognizes individual pairs (i,j) of a correct signal map, we explored how unique the resulting optimal (or near optimal) alignments are.

To evaluate this aspect, we compared two RPLC runs used before: T0SCX23A and T0SCX24A, representing successive SCX fractions of the yeast whole-cell lysate experiment, which can thus be expected to have some overlap in terms of their peptide content. Fig. 5 represents, for all spectra pairs (M_i,N_j) where M_i is the spectrum from T0SCX23A and N_j is the spectrum from T0SCX24A ($M_{i,j}$), the score of the best alignment containing (M_i,N_j) with *bright red* corresponds to the highest score. The *bright red* color around the main *diagonal* in the *highlighted* area represents the score of an optimal alignment, and the sharp change of colors around it indicates its robustness.

Note that, in the areas in which one of M_i or N_j contains no significant signal (here marked "*ND*"), an optimal alignment is not well defined as one would expect. The local "warps" are a well known phenomenon that can be exposed very clearly using this methodology, which renders signal maps as a useful tool in the refinement and validation of separation technology.

Optimal Signal Maps Are Correct—In the yeast data, after acquiring a mass spectrum up to four precursor masses were selected from that spectrum for CID. The selection was based on raw intensity levels, *i.e.* the higher peaks were preferentially selected for MS/MS. Thus, an MS spectrum was followed by zero to four MS/MS spectra. We used these alternately acquired

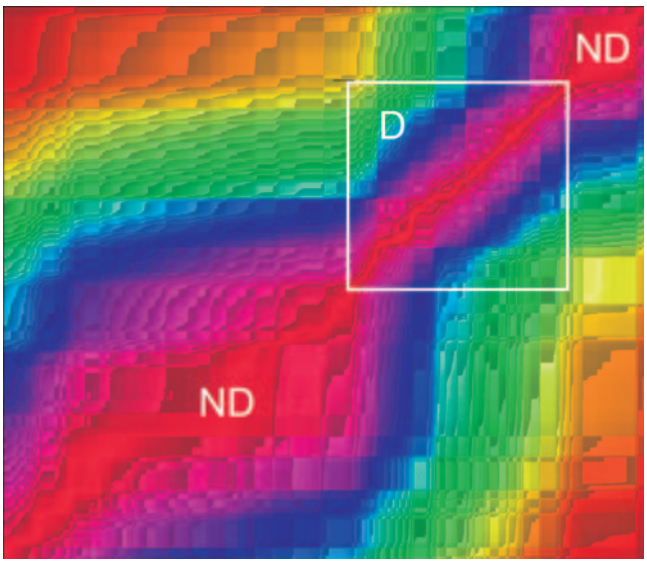

FIG. 5. **Image showing the robustness of the alignment path for two runs representing successive SCX fractions in the yeast whole-cell lysate experiment.** Indices *i* and *j* label the *horizontal* and *vertical axes*, respectively, where M_i and N_j are mass spectra of runs T0SCX23A and T0SCX24A, respectively. For all pairs (M_i,N_j), the score of the best alignment containing (M_i,N_j) is plotted using the color code shown on the *right*. The defined alignment region (*inset D*) shows the well defined best alignment (signal map) between the two runs. *ND*, no defined alignment.

CID spectra to evaluate the quality of the signal map.

As described earlier, experiments were performed at five time points in the yeast cell cycle with a significant degree of overlap expected between successive time points. For each time point a large number of CID spectra were conclusively assigned to a peptide ("identified") using Sequest (20) and PeptideProphet (21). As there is no experimental deterministic control for the elution phase in which any given peptide is sampled by CID, the correct signal map usually does not contain the corresponding MS spectra (k,l) but a pair (i,j) such that k is close to i and i is close to j.

We used the identified CID spectra in successive time points to assess the correctness of our signal map. Generally we found the successive time points to contain 10,000–20,000 identified CID spectra from shared peptides. The first step in our evaluation was to remove all the peptide identification information from time point 0. Then we transferred the peptide identifications from time point 30 to time point 0 using two algorithms. As described earlier in "Materials and Methods," the first approach is based entirely on spectra similarity, and the other approach puts additional constraints from the alignment (the two MS spectra from which the precursor masses were selected for CID are close according to the signal map). These transferred identifications can be compared with the original identifications of the various CID spectra of time point 0, thus computing the numbers of correct and incorrect transfers. Fig. 6 plots these numbers as we vary the threshold for the peptide transfer algorithm.

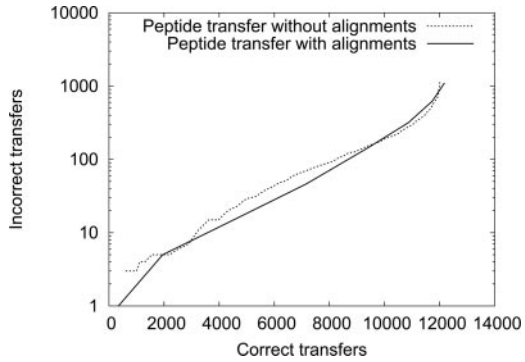

FIG. 6. **Correct peptide ID transfer rate against incorrect peptide transfer rate while transferring identifications from time point 30 to time point 0 in yeast whole-cell lysate experiment.**

As can be seen from the low rate of incorrect transfers for these plots in Fig. 6, our peptide ID transfer approaches are highly specific in transferring the correct identifications. Also the two plots are very similar to each other. If our signal map was bringing the wrong pair of spectra together, we would expect to see an increase in the false positive rate, and if it was not bringing the right pairs together, we will see a decrease in the true positive rate. These observations suggest that the signal map is correct. Similar inferences were made from analyses involving other pairs of time points.

Also using a very stringent threshold for the peptide transfer algorithm, we were able to transfer identifications to more than a thousand spectra of time point 0 that are not identified by Sequest. Thus this process of transferring identifications increases throughput and also represents a data set that can be used to understand the shortcomings of current approaches to identify peptides from mass spectra using database searches.

Signal Maps Can Identify Biomarkers—Signal maps can be used in various ways, but in this study we focused on one of their applications, *biomarker discovery*. Good biomarkers are signals whose presence (or absence) reliably indicates a biological state such as an early form of a health condition, two different growth conditions in cell cultures, etc. that would otherwise be hard to detect (22). Typically two collections of samples are compared in an attempt to identify the differences. The first collection of *case* or *disease* samples is from subjects who have a disease condition. The second collection of *control* samples is from subjects who do not have the condition. Biomarker discovery is the process of identifying signals that distinguish two such collections.

We applied signal maps to a synthetic biomarker discovery scenario in which the composition of all samples was known *a priori*. Our aim was to create a data set on which we would be able to analyze the performance of an experimental-computational approach to identify biomarker signals. Two protein samples were prepared, one "control sample" consisting of a four-protein digest and a second "disease" sample in which in addition to the four digested proteins β-lactoglobulin (as a simulated "biomarker") was spiked in. Six LC-MS experiments of the same four-protein control sample and seven experiments of the same five-protein disease sample formed the basis of our analysis. Despite having only four/five proteins, these mixture are surprisingly very complex as the peptides exhibit multiple charge and isotopic states, the proteins are not 100% pure, the tryptic digestion is imperfect leading to missed cleavages and miscleavages, and the gradient is short. All these issues arise in any proteomic setup increasing the complexity of the sample manyfold. This was observed for this mixture too as thousands of peptide-like features were observed in the experiments (without deisotoping), whereas the theoretical tryptic digestion yields only a hundred or so peptides.

The diagram in Fig. 7 summarizes our two-stage biomarker discovery strategy on the basis of signal maps. In the first stage, we created signal maps between the six runs (repeats) on the four-protein control sample, allowing us to analyze signals that occur consistently across theses runs and summarize them in a new *virtual* control run we called *C* (AnSi;

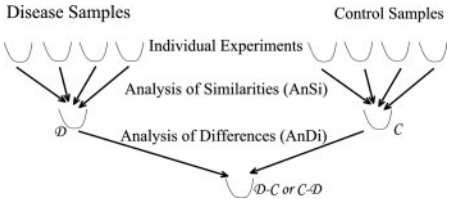

FIG. 7. **The two-stage biomarker discovery strategy based on signals maps.**

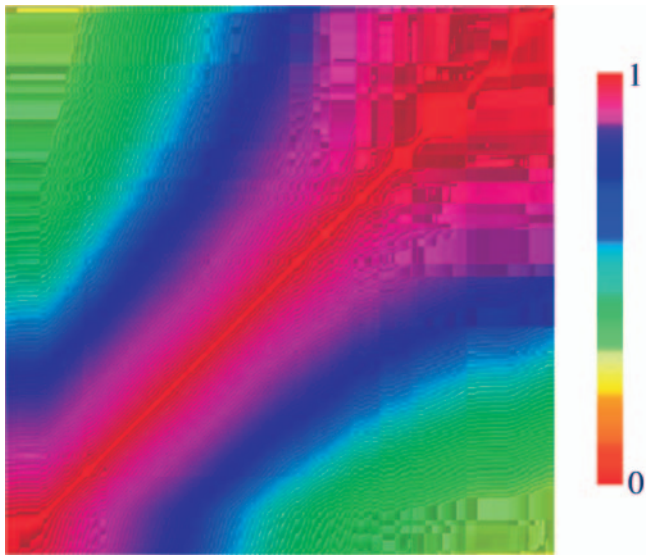

FIG. 8. **Image showing the alignment of two runs representing two repeats of the four-protein experiment.** Comparing this alignment to that of Fig. 5, we can see what a perfect alignment looks like. Indices *i* and *j* label the *horizontal* and *vertical axes*, respectively, where M_i and N_j are mass spectra of runs *R* and *S* (the two repeats). For all pairs (M_i,N_j), the score of the best alignment containing (M_i,N_j) is plotted using the color code shown on the *right*. The *bright red diagonal* region shows the well defined best alignment (signal map) between the two runs.

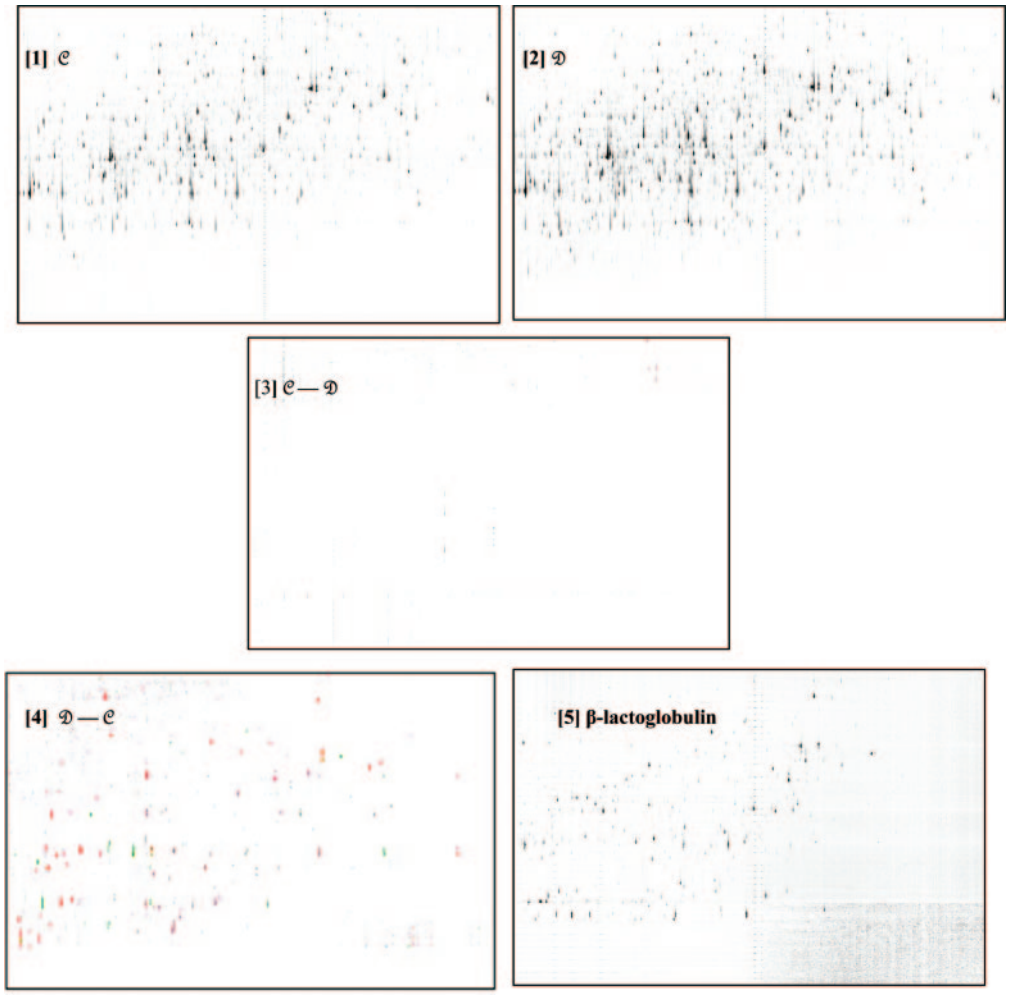

FIG. 9. **The virtual runs *C* (*panel 1*), *D* (*panel 2*) *C–D* (*panel 3*), *D–C* (*panel 4*), and β-lactoglobulin (*panel 5*) where *C* refers to the *control* run and *D* refers to the *disease* run.** For each run, *horizontal* and *vertical axes* correspond to *m/z* and elution order, respectively. Peak intensity is shown in *grayscale* with the most intense peak being plotted as *black*. Features detected computationally in the runs *C–D* (*panel 3*) and *D–C* (*panel 4*) are show in *red*. The *green* features in *D–C* (*panel 4*) are the ones that matched the features obtained from theoretical tryptic digestion of β-lactoglobulin. The significant overlap of *panels 4* and *5* suggests that our method is highly sensitive in identifying biomarkers, even the low intensity ones.

described earlier in "Materials and Methods"). This was done using pairwise alignments between the various repeat pairs. One such alignment is shown in Fig. 8. Comparing it with Fig. 5, we can see that the repeats were more reproducible in the four-protein mixture than in the yeast experiment.

The run *C* is shown in Fig. 9, *panel 1*, where horizontal and vertical axes correspond to *m/z* and elution order, respectively, and signal strength is represented by *gray* level. We then also performed the same process on the seven MS runs of the five-protein simulated disease sample, which yielded a virtual disease run (here denoted by *D*). As can be seen from Fig. 9, *panel 2*, and as expected, *D* contains more signals than *C*.

In the second stage, we constructed a signal map between *D* and *C* that allowed us to integrate these two data sets into a *virtual difference* run (using *AnDi*; described earlier in "Materials and Methods"). Briefly AnDi aims to identify signals present in the virtual five-protein run *D* but not in the virtual four-protein run *C*. Conveniently we call the resulting virtual run *D–C*. Before computing *D–C*, however, we performed a first validation of this approach, by computing *C–D*, a virtual run that should, under ideal experimental conditions *and* perfect signal maps, contain no signal. Fig. 9, *panel 3*, shows that indeed very few signals are present in *C–D*. In fact, there are hardly any signals present if we ignore the start and the end of the run *C–D*. This result confirms our assumption about peptide elution order: if the order in which any pair (*p*,*q*) of peptides elute had changed between the runs *C* and *D*, our alignment would have failed to capture that, and thus most likely signals from *p* or *q* (or both) would be present in *C–D*.

Fig. 9, *panel 4*, shows *D–C*, which contains many more signals, as expected under the correct signal maps. Colors in

the above figures were used to highlight those signals that an *ad hoc* feature detection method identified as features (described earlier in "Materials and Methods"). *Green* color was further used to highlight those signals that corresponded with a list of expected signals from β-lactoglobulin according to the theoretical tryptic digestion performed. The many features colored in *green* provide evidence that signal maps and feature recognition work.

The many features in *red*, however, represent features that are unexplained by our theoretical understanding of the experimental-computational process. We considered two likely possible sources of this phenomenon. On the one hand, faulty signal maps or faulty *AnSi-AnDi* analysis could well lead to the observed excess of signal in *D–C*, although the lack of signals in *C–D* has already suggested that AnSi-AnDi analysis is working. As a second possible source, the experimental protocol could have led to unexpected, non-tryptic cleavages in the fifth protein that would then be reflected as signals in *D–C*. Alternatively "pure" proteins are not always pure, and it might well be that there are contaminants in the five-protein sample. In addition there might be non-peptidic contaminants that might show up.

In an attempt to settle this issue, we performed a LC-MS/MS experiment with a sample that contained only a tryptic digest of β-lactoglobulin, the fifth protein, shown in Fig. 9, *panel 5*. On visual inspection, the figure displays a significant overlap with *D–C*, suggesting that the unexplained signals in *D–C* arise indeed from non-tryptic cleavages of the fifth protein or other signals arising only from the fifth protein. Conversely this means that AnSi-AnDi analysis on the basis of signal maps indeed reflects consistent differences between sets of experiments. Follow-up analysis revealed 14 non-tryptic peptides in MS/MS experiment of β-lactoglobulin, many of which overlap with the *red* signals shown in Fig. 9, *panel 4*. To quantify the overlap between *panels 4* and *5* of Fig. 9, we computed the virtual run ((*D–C*)- (β-lactoglobulin)) using the AnDi approach. The total ion current in this run was 15% of the total ion current in (*D–C*). This shows that *panels 4* and *5* of Fig. 9 overlap significantly in terms of the peptide content.

DISCUSSION

In this work we present algorithms to construct optimal signal maps between MS experiments and a practical application where these can be used to increase sensitivity and throughput. The approach relies on signal maps that associate raw data between experiments and the deferral of individual feature recognition to the last analysis stage. This approach can be expected to decrease inter- and intraexperiment biases and improve the signal-to-noise ratio, thus improving sensitivity and throughput. We present some preliminary results for the above claims where we show a significant overlap between the features identified by the signal maps as biomarkers (Fig. 9, *panel 4*) and the features experimentally seen as biomarkers (Fig. 9, *panel 5*).

Other potential uses of signal maps include the following.

- A peptide signal for which no CID has been performed in run *R* may be mapped to a peptide signal in run *S* in which CID was performed and identification exists. In many cases, this allows the purely computational identification of the peptide in run *R* without any CID. Preliminary evidence for this was shown in the yeast cell lysate experiment where we were able to transfer identification to nearly 10,000 CID spectra in time point 0 that were identified by Sequest and an additional 1,000 spectra that were unidentified by Sequest.
- Peptides that generate low intensity signals indistinguishable from noise in single runs may be detectable after the signals have been integrated across many runs, thus enhancing feature recognition methods.
- In the biomarker discovery application, we have presented ideas to identify discriminant features. The signal maps bring together signals generated from the same peptides in multiple runs. Thus they can even be used for identifying features common to these runs. Preliminary evidence for this was presented in the yeast cell lysate experiment where protein identifications were transferred from one time point to the other. Broadly speaking, signal maps can be used for quantitative comparisons of peptides/proteins across multiple runs.
- The signal map can be used to assess the reproducibility of various experimental designs, *e.g.* clinical trials of drugs, column optimization, etc. We can test the effects of the various experimental protocols on the quality of the alignment and thus differentiate variations among technical and biological replicates.
- The idea of a signal map extends beyond aligning runs. Instead we can think of aligning SCX fractions and time points as even successive SCX fractions and time points have similarity in their peptide contents. We are currently designing methods to do this as alignments at these levels have the potential for further enhancements of sensitivity and specificity of peptide detection.

All computational analyses were performed on Linux personal computers. Computing the optimal signal map between two runs (1.5-gigabyte mzXML (23) files each) takes around 10–20 min on a normal work station. We are currently working to develop a better understanding of elution curves to develop statistical approaches for the feature detection method that follows the AnDi analysis.

Although we have presented evidence that alignment can lead to the cited benefits on low-to-moderate resolution mass spectrometers, other platforms may require different parameter settings. We believe that, despite these foreseeable adaptations, we have shown that the approach is quite powerful. In particular, most of the data we used here were acquired on instruments with low-to-moderate resolution and accuracy, such as LCQ, Q-TOF, and LTQ. It will be very interesting to run our tools on data sets acquired on high resolution and more accurate instruments, such as FTICR.

Acknowledgments—Many of our colleagues at the Institute for Systems Biology and Fred Hutchinson Cancer Research Center contributed to this work. In particular we thank Jimmy Eng, Eric Deutsch,

Brian Piening, Martin McIntosh, and the anonymous reviewers, who suggested the new notation we use here for sets of experiments and their parts.

* This work was supported in part by federal funds from the NHLBI, National Institutes of Health under Contract Number N01-HV-28179. 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.

¶¶ To whom correspondence should be addressed. Tel.: 33-1-4568-8620; Fax: 33-1-4061-3704; E-mail: benno@pasteur.fr.

REFERENCES

1. Ideker, T., Galitski, T., and Hood, L. (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2,** 343–372
2. Aebersold, R., and Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chem. Rev.* **101,** 269–295
3. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
4. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50
5. Tyers, M., and Mann, M. (2003) From genomics to proteomics. *Nature* **422,** 193–197
6. Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P., King, N. L., Eng, J. K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Hood, L., Katze, M. G., Kennedy, K. A., Kregenow, F., Lee, H., Lin, B., Martin, D., Ranish, J. A., Rawlings, D. J., Samelson, L. E., Shiio, Y., Watts, J. D., Wollscheid, B., Wright, M. E., Yan, W., Yang, L., Yi, E. C., Zhang, H., and Aebersold, R. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6,** R9
7. Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y., Conrads, T. P., Veenstra, T. D., and Udseth, H. R. (2002) An accurate mass tag strategy for quantitative and high throughput proteome measurements. *Proteomics* **2,** 513–523
8. Beer, I., Barnea, E., Ziv, T., and Admon, A. (2004) Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **4,** 950–960
9. 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
10. Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T. G., Foss, E., Mao, Y., and Emili, A. (2004) Informatics platform for global proteomic profiling and biomarker discovery using liquid-chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **3,** 984–997
11. Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **75,** 4818–4826
12. Snyder, L. R., Kirkland, J. J., and Glaich, J. L. (1997) *Practical HPLC Method Development*, 2nd Ed., pp. 214–277, Wiley Interscience, New York
13. Bylund, D., Danielsson, R., Malmquist, G., and Markides, K. E. (2002) Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography mass spectrometry data. *J. Chromatogr. A* **961,** 237–244
14. Listgarten, J., Neal, Radford M., Roweis, Sam T., and Emili, A. (2005) Multiple alignment of continuous time series, in *Advances in Neural Information Processing Systems*, Vol. 17 (Saul, L. K., Weiss, Y., and Bottou, L., eds.) pp. 817–824, MIT Press, Cambridge, MA
15. Miller, W., Makova, K. D., Nekrutenko, A., and Hardison, R. C. (2004) Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5,** 15–56
16. Sakoe, H., and Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust.* **ASSP-26,** 43–49
17. Stein, S. E., and Scott, D. R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5,** 859–866
18. Needleman, S. B., and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453
19. Kruskal, J. J. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.* **7,** 48–50
20. Eng, J., McCormack, A., and Yates, J. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass. Spectrom.* **5,** 976–989
21. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of protein identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
22. Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359,** 572–577
23. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22,** 1459–1466