research articles Journal of proteome research

Added Value for Tandem Mass Spectrometry Shotgun Proteomics Data Validation through Isoelectric Focusing of Peptides

Manfred Heller,[†,‡] Mingliang Ye,[§] Philippe E. Michel,[†] Patrick Morier,[†] Daniel Stalder,[‡] Martin A. Jünger,[∥] Ruedi Aebersold,[§,∥] Frédéric Reymond,[†] and Joël S. Rossier*[,†]

DiagnoSwiss SA, Monthey, Switzerland, Department of Clinical Research, University Hospital, 3010 Bern, Switzerland, Institute for Systems Biology, Seattle, Washington 98103, and Institute for Molecular Systems Biology, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland.

Received June 24, 2005

A very popular approach in proteomics is the so-called "shotgun LC−MS/MS" strategy. In its mostly used form, a total protein digest is separated by ion exchange fractionation in the first dimension followed by off- or on-line RP LC−MS/MS. We replaced the first dimension by isoelectric focusing in the liquid phase using the Off-Gel device producing 15 fractions. As peptides are separated by their isoelectric point in the first dimension and hydrophobicity in the second, those experimentally derived parameters (p*I* and R_T) can be used for the validation of potentially identified peptides. We applied this strategy to a cellular extract of *Drosophila* Kc167 cells and identified peptides with two different database search engines, namely PHENYX and SEQUEST, with PeptideProphet validation of the SEQUEST results. PHENYX returned 7582 potential peptide identifications and SEQUEST 7629. The SEQUEST results were reduced to 2006 identifications by validation with PeptideProphet. Validation of the PeptideProphet, SEQUEST and PHENYX results by p*I* and R_T parameters confirmed 1837 PeptideProphet identifications while in the remainder of the SEQUEST results another 1130 peptides were found to be likely hits. The validation on PHENYX resulted in the fixation of a solid *p*-value threshold of $<1 \times 10^{-04}$ that sets by itself the correct identification confidence to >95%, and a final count of 2034 highly confident peptide identifications was achieved after p*I* and R_T validation. Although the PeptideProphet and PHENYX datasets have a very high confidence the overlap of common identifications was only at 79.4%, to be explained by the fact that data interpretation was done searching different protein databases with two search engines of different algorithms. The approach used in this study allowed for an automated and improved data validation process for shotgun proteomics projects producing MS/MS peptide identification results of very high confidence.

Keywords: LC−MS/MS • isoelectric focusing • retention time • peptide identification • database searching • proteomics • data validation

Introduction

The term proteomics was coined in 1994 to describe a technology with which the entire protein complement of a biological sample could be described.[1] At that time, proteomics was based on the technique of two-dimensional gel electrophoresis separating proteins in the first dimension by ioselectric point followed by molecular weight in the second dimension. This technique seemed to be suitable for the separation of several thousand proteins including isoforms because on a large format gel several thousand spots could be detected. With performance improvements achieved in mass spectrometry during the second half of the past decade more and more 2D gel spots could be submitted for protein identification. Soon it was realized that despite the fact that 2D gels offer a tremendous separation power only few, and mostly high abundant proteins were detectable and identifiable with 2D gels. This restriction can be explained by a limited loading capacity, tremendous differences in physicochemical properties of proteins, problems of transferring proteins from the first dimensional gel onto the second one, and artifacts introduced during sample treatment.[2] Technical advances on several fronts spurred the development of alternative proteomics methods with a trend away from labor intensive gel technology to more automatable methods encompassing a more direct hyphenation of mass spectrometry with liquid chromatography and working with proteolytic digests of complex protein samples rather than intact proteins in order to reduce physicochemical constraints.[3,4] Both approaches, the 2-dimensional peptide

* To whom correspondence should be addressed. Dr. Joel S. Rossier, DiagnoSwiss SA, Route de l'Ile-au-Bois 2, c/o CIMO SA, CH-1870 Monthey, Switzerland. Fax: +41 24 471.49.01. E-mail: j.rossier@diagnoswiss.com.

[†] DiagnoSwiss SA.

[‡] Department of Clinical Research, University Hospital.

[§] Institute for Systems Biology.

[∥] Institute for Molecular Systems Biology, Swiss Federal Institute of Technology (ETH).

10.1021/pr050193v CCC: $30.25

Published on Web 10/05/2005

separation by strong cation exchange and reversed phase liquid chromatography and the isotope coded affinity technology have since been refined and combined for increased proteome coverage.[5] More recently, different groups separated intact proteins in the liquid phase with multidimensional chromatography before mass spectrometric detection and identification.[6−9] A very interesting attempt that was applied to subproteomes is the combination of a top-down and bottom-up approach. The exact mass of intact proteins was determined for the detection of potential isoforms and the relative quantification of the different proteins based on the mass spectrometry signal, and identification of proteins by mass spectrometric analysis of the protein digests.[10,11] Another tool for fractionation of proteins and peptides based on their isoelectric point called off-gel isoelectric focusing (OGE) has been recently introduced and presented as a versatile device.[12,13] Intact human plasma protein isoforms differing in p*I* were separated in a first stage IEF followed by proteolysis of the proteins and subsequent isoelectric focusing of the peptides with the same device. The resulting 225 peptide fractions (15 protein fractions and 15 peptide fractions of each protein fraction) were directly amenable for LC−MS/MS analysis. Such large scale proteome analyses result in a tremendous amount of peptide identification data. Validation of these data becomes a major task and currently researchers rely on peptide and/or protein scoring values returned by the identification software. Doing so, one has to confront identification data that is either contaminated with a high percentage of false positive identifications or a big loss of correct identifications, termed false negatives, as illustratively documented by Cargile and colleagues.[14] This had been recognized by different groups who subsequently engineered new statistical models allowing for the validation of identification results returned by the database searching software, mostly SEQUEST.[15−17] More recently, the new identification software OLAV was developed and is now available as PHENYX from GeneBio.[18] While the frequently used SEQUEST [19] and MASCOT [20] software identify peptides by correlation of experimental with theoretical fragmentation spectra without involving a model, PHENYX uses a more efficient scoring scheme based on signal detection theory coupled with pattern recognition for likelihood ratio calculation in order to distinguish correct from false peptide identifications. In addition to improvements on the software side, the incorporation of experimentally measured peptide parameters into the peptide identification process will increase significantly the correct identification ratio in shotgun proteomics.

It was the group of Stephenson that has suggested with a couple of publications in 2004 to use experimentally determined peptide p*I* values for validation of the correctness of peptide identifications in shotgun proteomics experiments.[14,21] We have successfully used this concept for the human plasma protein work published recently and have combined it in a manual fashion with the relative retention time behavior of peptides from the reversed phase column.[13] The calculation of peptide retention times was first reported by Martin in 1948[22] and since then different groups published peptide retention time prediction algorithms. One of the more elaborate algorithm was developed by Sakamoto et al.,[23] and more recently, the group of Smith has refined and used it as a peptide identification criteria.[24,25] We have written a Microsoft Excel macro that computes pH values, charge states at defined p*I*'s, and a predictive retention time for peptides identified by LC−MS/MS. Here, we report the application of these tools for the confident peptide identification from a tryptic digest of a total *drosophila* cell lysate isoelectrically fractionated on the OGE device, pH 3−10, into 15 fractions with subsequent LC−MS/MS analysis of each OGE fraction. The MS/MS data was searched with SEQUEST and PHENYX against relevant protein databases. The raw identification results were treated with the excel macro and by applying acceptance criteria for p*I* and retention time the raw data was filtered and cleaned from false positive identifications. The comparison of the identification results returned by both search engines was further used to validate this approach and establish acceptance criteria for peptide identifications with PHENYX.

Materials and Methods

Sample Preparation for OGE Fractionation. The Drosophila Kc167 cell line cell line was generously provided by E. Hafen (Zoological Institute, University of Zurich) and was originally derived from disaggregated 8 to 12 h old *Drosophila* embryos.[26] Cells were cultured in Schneider medium containing 10% FCS at 25 °C. Density-arrested Kc167 cells were washed with PBS, collected in hypotonic lysis buffer (10 mM HEPES−KOH pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 1x Complete Protease inhibitor cocktail, Roche) and lysed by Dounce homogenization. After centrifugation at 100 000 × *g* for 30 min, the pellet was discarded, and cytoplasmic proteins were precipitated with acetone. After precipitation, proteins were redissolved in 100 mM Tris-HCl buffer (pH 8.3) including 0.05% SDS, 5 mM EDTA and 6 M urea. An aliquot of 2 mg protein was reduced by 5 mM tributylphosphine (Aldrich) and alkylated by 10 mM iodoacetamide. The sample was then 8-fold diluted with water and sequencing-grade modified trypsin was added at an enzyme: protein ratio of 1:50 w/w and incubated at 37 °C overnight. The *Drosophila* protein digest was desalted by strong cation exchange (SCX) purification as follows. The digest solution was acidified to pH 2.8 by 1% formic acid, and then loaded onto a SCX SPE cartridge (Polysulfoethyl Aspartamide, PolyLC, Columbia, MD). After extensive washing with 4 mL of 0.1% (v/v) formic acid and 2 mL of 25% (v/v) acetonitrile in 0.1% (v/v) formic acid peptides were eluted with 1.5 mL of 10% (v/v) NH_4OH solution containing 25% (v/v) acetonitrile. The purified peptides were evaporated to dryness in a vacuum centrifuge and redissolved in OGE buffer composed of 5% (v/v) glycerol, 0.5% (v/v) ampholytes pH 3.0−10.0 (Amersham Biosciences, Otelfingen, Switzerland) in pure water.

OGE Fractionation of Peptides. Isoelectric focusing of peptides was performed with the off-gel electrophoresis device (OGE) composed of 15 wells over a 13 cm IPG strip (Amerham Biosciences, Otelfingen, Switzerland) exhibiting a linear pH gradient ranging from 3 to 10 as described elsewhere.[13] Briefly, the separations were run by dispensing 50 μL of peptide solution in each well (total of 750 μL) and the potential was fixed during 1 h at 500 V, then 1 h at 1000 V and finally 3.5 h at 8000 V (total of 29.5 kVh). The current limit was set at 200 μA per strip and the temperature was maintained at 20 °C. Fractionations were run with a peptide loading equivalent to 100 μg of the starting protein preparation. After OGE, liquid fractions were withdrawn (20 μL in average) and the OGE wells were rinsed once in order to enhance the peptide yield. For this purpose 100 μL of a water/methanol/formic acid (49:50:1 by volume) mixture was added per well and incubated for 90 min without voltage. Corresponding peptide fractions from 10 runs were pooled and concentrated by vacuum centrifugation. Each OGE peptide fraction was purified from residual traces

of glycerol, urea and ampholytes as follows. The fractions were diluted to 1 mL in 0.1% (v/v) TFA and acidified to pH 3.0 by addition of 1% (v/v) TFA, then loaded onto a Sep-Pak C18 cartridge (Waters) for purification as recommended by the manufacturer. The purified fractions were then evaporated to dryness and redissolved in 20 μL of 0.4% (v/v) acetic acid for capillary RPLC−MS/MS analysis.

RPLC−MS/MS Analysis. The setup of the capillary RP−LC system was as described previously.[27] The system consisted of a binary HPLC pump (HP1100, Agilent Technologies, Wilmington, DE), a micro-autosampler (Famos, Dionex LC Packings, San Francisco, CA), a ten-port switching valve integrated on a Finnigan ion trap mass spectrometer (model LCQ XP, Thermo Electron Corporation, San Jose, CA), a precolumn (100 μm i.d. $\times$ 2.0 cm length), and an analytical capillary column (75 μm $\times$ 12 cm). Fused silica capillary tubing with an integrated borosilicate frit (Integrafrit, New Objective, Cambridge, MA) was used for the precolumn. For the capillary column, one end of polyimide-coated fused-silica capillary (Polymicro Technologies, Phoenix, AZ) was manually pulled to a fine point ~5 μm with a micro-flame torch. The columns were in-house packed with C18 resin (5 μm, 200 Å Magic C18AQ, Michrom BioResources, Auburn, CA) using a pneumatic pump (Brechbuehler, Spring, TX) at constant helium gas pressure of 1500 psi. Sample volumes of 6 μL were loaded onto the precolumn at a flow rate of 5 μL/min in 5 min with solvent A {0.1% (v/v) formic acid in water}. After sample loading and cleanup, a linear binary gradient of 5−35% solvent B {0.1% (v/v) formic acid in acetonitrile} in solvent A over 80 min was applied, followed by isocratic elution at 80% B for 10 min. Peptides eluting from the capillary column were selected for CID by the mass spectrometer using a protocol that alternated between one MS scan and three MS/MS scans.

Data Processing. MS/MS spectra were interpreted by SEQUEST from a *drosophila* sequence database downloaded from NCI (ftp://ftp.ncifcrf.gov/pub/nonredun/protein.nrdb.Z) at the Institute for Systems Biology in Seattle. Carbamidomethylated cystein was set as a fixed modification and oxidation of methionin as a variable modification. Furthermore, no enzyme specificity was chosen and a mass difference of $\pm$ 3 Da and $\pm$ 0.5 Da for precursor and fragment ions, respectively, was accepted. The database search results were validated using the PeptideProphet program.[15] PeptideProphet assigns to each SEQUEST peptide identification a probability that it has been correctly identified based upon its SEQUEST scores and additional information of the assigned peptides, including the number of tryptic termini. Peptides with a probability of 0.9 or higher were considered a match in this study. Peptide identification by PHENYX was performed on the vital-it processor cluster at the EPFL in Lausanne, Switzerland (www.phenyx-ms.com/). The Uniprot-Tremble database release 26.2 and SwissProt release 46.0 were searched with restriction to *Drosophila melanogaster* protein entries only. Otherwise, the same search criteria were applied as for the SEQUEST search except that PHENYX requests a protease specificity (trypsin) but allowing for half-tryptic peptides with up to two missed cleavages. Furthermore, only PHENYX identifications with a *p*-value of <1 were accepted.

The complete lists of identified peptides were extracted into Microsoft excel.csv files by a program written in Perl at the University of Bern. The large lists of redundant peptide identifications were treated as follows. In case of PHENYX data, only the identification with the smallest *p*-value of each precursor ion fragment spectrum was kept. Such a filtering strategy is a valid one because the PHENYX *p*-value is a statistical measure of the probability that the identification is a random match (note: the bigger the value the bigger this probability).[18] In the case of SEQUEST results, the sorting was performed according to parameters that have been widely applied in the existing shotgun proteomics literature (for example Cargile and colleagues).[14] In a first round, only peptides that had at least one trypsin-specific cleavage site were accepted. If there were still two identifications on the same MS/MS spectrum, then the following hierarchy of acceptance decisions was followed through: First, the identification on a fully tryptic peptide was preferred over a half tryptic peptide. Second, peptides with Xcorr scores larger than 1.5 in case of singly charged, 2.0 for doubly charged, and 2.2 for triply charged peptides, respectively, were preferred. Third, the peptide with the highest dCn score was finally preferred. This strategy was validated by the observation that PeptideProphet, a recently developed computing tool based on empirical statistical modeling to predict correct peptide assignments,[15] assigned 2006 out of the 34 118 SEQUEST peptide identifications with a probability of 0.9 and bigger, while 2000 of those identifications were present in our sorted, nonredundant (nr) SEQUEST list. Furthermore, we reduced each identification list by keeping only the best scoring peptide compound from each OGE fraction.

A visual-basic macro available at DiagnoSwiss SA (info@diagnoswiss.com) was used to calculate theoretical p*I*'s and predictive RP retention times of all peptides according to computing algorithms proposed by Bjellquist et al [28] and Guo et al.,[29] respectively. The visual-basic macro allows the user to compute peptide p*I*'s with amino acid values proposed by many different authors. Here, we calculated all p*I*'s with the values given in the Lehninger biochemistry text book.[30] It also respects deamidation of asparagine and glutamine to the corresponding acids that renders a peptide p*I* more acidic. However, we have not included this feature in the current data validation because SEQUEST output files did not distinguish between a better match on a deamidated or native peptide. The validation of peptide identifications was performed as outlined in detail in the Results section below using Microsoft excel and Matlab software.

Results

OGE was first introduced for the isoelectric separation and recovery of proteins in solution and its application was extended to the separation of peptides.[12,13] We have shown its excellent performance and reproducibility on the fractionation of intact proteins and peptides of human plasma. The value of integrating the experimentally derived peptide parameters p*I* (OGE) and retention time (R_T) of the reversed phase chromatographic separation (LC−MS/MS) into the validation of peptide identification results was already touched in our earlier report.[13] Here, we report the evaluation of this idea by using a Visual Basic macro for the automatic calculation of peptide p*I* and approximated R_T together with peptide scores from peptide identification software. Comparison of SEQUEST and PHENYX results was then used to validate this approach.

Data Assessment. An aliquot corresponding to 1 mg of a soluble protein extract from the *drosophila* cell line Kc167 was digested with trypsin and the peptides were subsequently separated with OGE into 15 fractions according to isoelectric point followed by LC-MS/MS analysis of an aliquot corre-

Table 1. Experimentally Derived and Theoretically Calculated p*I* Values in Each of the 15 OGE Fractions

OGE Fr. no.	theoret. range[a]	calcd. median	calcd. mean	Stdev	narrow range[b]	wide range[c]
1	3.70−4.02	3.69	3.74	0.50	3.22−4.21	2.73−4.71
2	4.08−4.40	4.08	4.05	0.25	3.76−4.27	3.51−4.53
3	4.45−4.78	4.45	4.45	0.58	3.89−5.06	3.30−5.64
4	4.83−5.15	4.63	4.97	0.95	4.08−5.97	3.14−6.92
5	5.21−5.53	5.22	5.61	0.92	4.74−6.59	3.82−7.51
6	5.58−5.91	6.29	6.05	0.85	5.30−7.01	4.45−7.86
7	5.96−6.28	6.83	6.63	0.62	6.13−7.37	5.51−7.99
8	6.34−6.66	6.87	6.69	0.70	6.09−7.48	5.39−8.18
9	6.71−7.04	7.74	7.23	0.84	6.57−8.25	5.73−9.09
10	7.09−7.41	7.76	7.51	0.41	7.24−8.07	6.83−8.48
11	7.47−7.79	7.82	7.65	0.49	7.21−8.20	6.72−8.69
12	7.84−8.16	7.84	8.27	1.21	7.08−9.50	5.88−10.70
13	8.22−8.54	9.63	9.41	0.72	8.76−10.21	8.04−10.94
14	8.60−8.92	9.72	9.46	0.62	9.04−10.27	8.42−10.89
15	8.97−9.29	9.85	9.70	0.68	9.15−10.50	8.47−11.18

[a] Theoretical pH range in each OGE well calculated on IPG strip specifications and OGE device geometry. [b] pH intervall calculated from the mean p*I* of identified peptides in each OGE fraction ± 1 Stdev. [c] pH intervall calculated from the mean p*I* of identified peptides in each OGE fraction ± 2 Stdev.

sponding to 30% of each OGE fraction. The 15 MS/MS peak lists were interrogated with SEQUEST and PHENYX as described under Materials and Methods. PHENYX returned a total of 23 622 and SEQUEST 34 118 peptide identifications, respectively. The large peptide identification lists contained redundant identifications because each fragmentation spectrum not recognized as a singly charged ion was exported twice as a doubly and triply charged precursor into the corresponding peak list by the data processing software and there were no or very low rejection thresholds used in order to not lose any potentially correct but low scoring identification. The redundant data was processed as described under Materials and Methods breaking the numbers down to a total of 7582 and 7629 peptides for PHENYX and SEQUEST, respectively (Tables 1, 2, and 3 in Supporting Information).

The theroretical p*I* of each peptide was calculated by the visual basic macro. In our previous report we have already shown that the mean or median p*I* calculated in each OGE fraction was not a uniform function of fraction number and did not exactly follow the theoretical pH range in each OGE chamber (Table 1).[13] PHENYX *p*-value and SEQUEST dCn values were plotted against all the calculated peptide p*I* of the nonredundant lists (Figure 1, panels A and B). Both *p*-value and dCn are a measure for the probability of a correct match. While the *p*-value is a real statistical entity for the probability of the correctness of the match the dCn is only a discriminator between the peptide scores of the best and the second best match. The distribution of p*I* values of the nr peptide lists revealed a clustering into distinct p*I* zones with a more distinct clustering the better the score values for PHENYX and SEQUEST identifications were. The same type of clustering was observed with 7400 randomly chosen tryptic peptides generated in silico from *Drosophila melanogaster* protein sequences identified by PHENYX from the Uniprot database in this study (Figure 1, panel D). The observed clustering is possibly a consequence of the way how peptide p*I* values are computed from distinct pK_a values of only a few charged amino acid residues making up the building blocks of peptides. Furthermore, the PeptideProphet validated SEQUEST results showed good clustering independently of the dCn score (Figure 1, panel C). Interestingly, both, PHENYX and SEQUEST, had

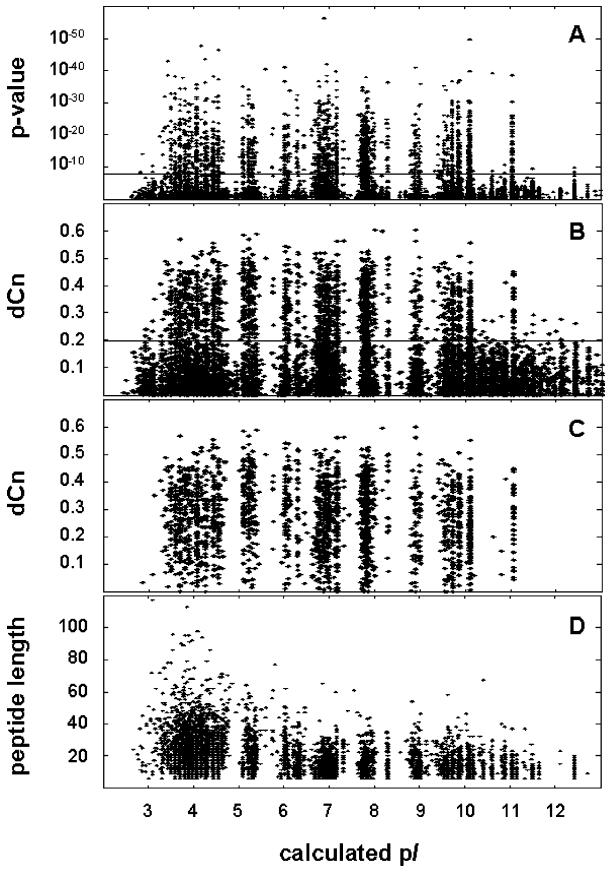

Figure 1. Peptide p*I* distribution of the soluble *Drosophila* proteome. The calculated p*I* values of identified peptides were plotted against a scoring value, that is *p*-value in panel A of the nonredundant PHENYX, dCn in panel B of the nonredundant SEQUEST, and dCn in panel C of the PeptideProphet validated SEQUEST identifications, respectively. The horizontal lines in panel A and B denote the score threshold value above which data was used for setting filtering criteria that is p-value $< 1 \times 10^{-07}$ and dCn ≥ 0.2, respectively. Panel D is a representation of the calculated peptide p*I* values of a random list of 7400 peptides taken from the PHENYX identified *drosophila* proteins after in silico digestion according to trypsin specificity allowing for up to two missed cleavage sites. The p*I* values were plotted against the length of peptides starting with a minimum of 6 amino acids.

many low scoring identifications at the very acidic and basic end of the p*I* spectrum that were not present in the PeptideProphet validated SEQUEST data. Otherwise the clusters were at the same p*I* regions in all data sets and occurred at very specific p*I* zones.

From our previous experience we are confident to have a very high percentage of correct PHENYX identifications with a p-value $< 1 \times 10^{-07}$ corroborated by the nicely aligned and reproducible p*I* clustering of peptide identifications fulfilling this *p*-value threshold as shown in Figure 1, panel A.[13] In case of SEQUEST a generally applied acceptance criteria found in the literature is a dCn = 0.1 cutoff value. With this cutoff value a greater proportion of possibly false identifications remained in the data set based on the observation of a less focused p*I* distribution compared with the PHENYX and PeptideProphet results. For this reason, only peptide identifications with a dCn ≥ 0.2 were used to plot the p*I* distribution in each OGE fraction (Figure 2). The average standard deviation (stdev) in this data set was 1.05 ± 0.30 p*I* units, compared with 2.06 ± 0.30 p*I* units

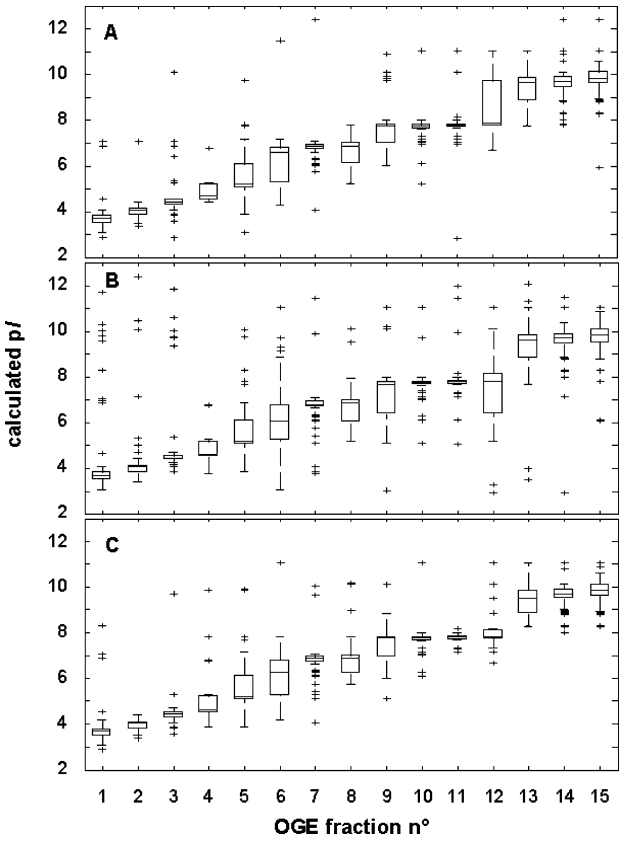

Figure 2. Box and whisker plot of calculated peptide p*I* in each of the 15 peptide OGE fractions. Peptides identified by PHENYX with a *p*-value $< 1 \times 10^{-07}$ were considered in panel A, peptides identified by SEQUEST with a dCn value of ≥ 0.2 in panel B, and SEQUEST peptides validated by PeptideProphet in panel C, respectively. The boxes represent the 25% to 75% with the median value as a horizontal line. The whiskers encompass the range of data points that can statistically be considered to belong to the data set with a 95% probability. Statistical outliers are marked by crosses.

with the dCn = 0.1 threshold. The p*I* distribution of the three datasets correlated relatively well as can be expected under the assumption that only highly certain peptide identifications were used (Figure 2). There were however a few differences. SEQUEST peptides were more heterogeneous reflected in a bigger standard deviation of calculated mean p*I* (see above) compared with 0.90 ± 0.22 p*I* units for PHENYX results. In addition, the p*I* distribution of PHENYX fraction 12 was skewed to higher values while the corresponding SEQUEST data was skewed toward more acidic values, as it was the case with most other fractions of both data sets (Figure 2). Consequently, we have chosen PHENYX identified peptides with *p*-value $< 1 \times 10^{-07}$ together with all PeptideProphet identifications in order to determine the p*I* and R_T acceptance criteria. The resulting median and mean p*I* values of the combined data of each OGE fraction are given in Table 1 together with the theoretically calculated p*I* range based on the well dimensions of the OGE device. The theoretical p*I* range in each fraction was generally well covered by the experimentally derived mean p*I* ± 1 stdev and always within the limits of the wide range with ±2 stdev.

Figure 3 is a graphical representation of the approximated R_T values in relation to the MS scan number of the same subset of peptide identifications. MS scan numbers were used as the experimental R_T value because the Finnigan LCQ software

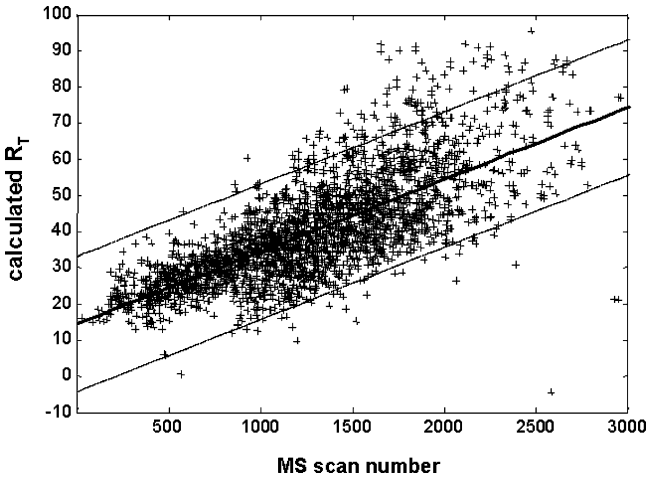

Figure 3. Linear MS scan number to R_T curve fitting with the combined PeptideProphet and high scoring PHENYX (*p*-value $< 10^{-7}$) peptide identifications. Linear curve fitting was done with the least-squares method applying a 95% confidence interval (thin lines) using Matlab (lower acceptance boundary: $y = 0.01991 * x - 4.203$; upper acceptance boundary: $y = 0.01991 * x + 33.175$).

exports only scan numbers into dta files and not effectively measured run time. As each MS scan on the LCQ iontrap lasts approximately the same period of time, MS scan numbers correlate well with actual run time. The individual linear curve fitting on PHENYX and PeptideProphet data resulted in similar trend lines with slopes of 0.0193 and 0.0204, correlations of 0.5554 and 0.5480, and intercepts of 14.651 and 14.572, respectively. For the subsequent data filtering we used the R_T acceptance criteria given in the legend of Figure 3 and the mean p*I* ± 1 or 2 stdev boundaries (narrow or wide boundary) as given in Table 1.

Data Filtering. From Figure 2, it becomes apparent that isoelectric focusing of peptides by OGE does not result in a linear trend of increasing p*I* from well to well. Hence, it is not absolutely clear what amount of variability can be allowed for as acceptance criteria of an identified peptide. For this reason, we tested additional criteria to affirm the proposed validation process. The chromatographic R_T behavior of peptides is orthogonal to isoelectric point and was consequently included in all data treatment with the relatively relaxed conditions given in Figure 3. We searched the *Drosophila* data against the nonrelevant Swiss-Prot human protein database in order to generate the false positive identification rate of the PHENYX search engine. The list of 33171 returned identifications, all with *p*-values < 1, consisted of mostly low scoring peptides with only 4.3% of *p*-values $< 1 \times 10^{-04}$. This list was treated as described above, resulting in 8102 nonredundant identifications (9% having *p*-value $< 1 \times 10^{-04}$), 170 of which were identical with sequences also found searching the *drosophila* protein database. A much higher fraction of 78.8% among this 170 identifications had a *p*-value $< 1 \times 10^{-04}$. After removal of the *Drosophila* peptides, R_T and p*I* filtering was performed using one or two p*I* stdev boundaries leaving 1064 or 1846 identifications in the list, respectively. It became obvious that both ways of filtering did not change significantly the outcome of the data quality as seen in Figure 4. This is explained by the fact that the identifications are random, false positive matches that cannot be selectively filtered by a criteria based on specificity. It is worth mentioning that the p*I* values of the human peptide

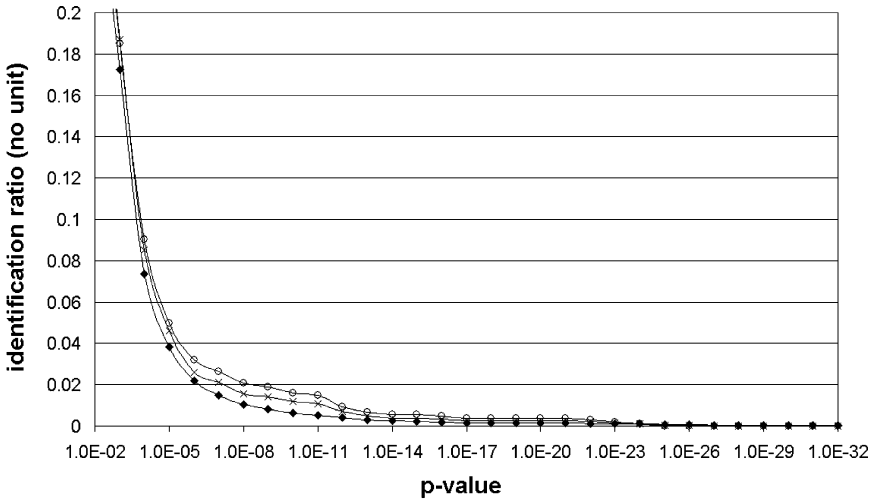

Figure 4. Establishment of false positive identification rates with PHENYX peptide identification engine. The diamonds represent the fraction of all identifications made on a human protein database with CID spectra from the *Drosophila* peptides with a *p*-value equal or smaller than the value given on the *x*-axes. Open circles and crosses stand for the R_T and p*I* filtered datasets applying ±1 or ±2 p*I* stdev tolerances, respectively.

identifications did not cluster as nicely as it was the case with the *Drosophila* identifications corroborating above statement (results not shown). In conclusion, the curves in Figure 4 show that less than 5% of PHENYX identifications from a nonrelevant database had a *p*-value of $< 1 \times 10^{-04}$, hence it can be concluded that a confidence level of equal or better than 95% was possible with this *p*-value threshold.

We applied the same validation strategy to all data sets of *Drosophila* peptide identifications returned by PHENYX, SEQUEST, and PeptideProphet software. The breakdown in numbers of identified peptides was as follows: Untreated nr PHENYX 7582, 3059 remained with a ±2 p*I* stdev interval, and 2286 fulfilled the ±1 p*I* stdev interval; SEQUEST filtering resulted in 7629, 2961, and 2140, and PeptideProphet in 2006, 1837, and 1528 identifications, respectively (Tables 1, 2, and 3 in Supporting Information). The distribution ratios according to score values of all the data are presented in Figure 5. It is obvious that the R_T and p*I* validation increased the ratio of better scoring peptide identifications with PHENYX and SEQUEST, and the narrower tolerance range of only ±1 p*I* stdev had a stronger effect than ±2 p*I* stdev. The overall quality of PeptideProphet validated SEQUEST results were not changed by R_T and p*I* filtering, and it can be concluded that PeptideProphet validated identifications can be considered as true positives. In addition, from the 169 PeptideProphet identifications that were removed by the R_T and wide p*I* filter only 60 did not fulfill the p*I* criteria. Seven of which, like all the other 108 identifications not accepted, were not accepted because they were outside the R_T acceptance range. Therefore, it was concluded that despite the fact that a ±1 p*I* stdev filter results in better data quality than the ±2 p*I* stdev filter there are most likely considerable losses of true positive identifications with a too narrow filter. Otherwise the PeptideProphet curves were close to the ideal case with close to 100% of identifications having a score above a certain confidentiality threshold. Both, the PHENYX and SEQUEST curves showed a linear increase in numbers of identified peptides from low *p*-values up to 1×10^{-05} and high dCn values down to about 0.08, respectively, thereafter the curves started to be irregular. Again, this illustrates that with higher *p*-values than 1×10^{-05} the false identification rate increases as illustrated in Figure 4. In conclusion, by bearing in mind that with the PHENYX *p*-value there is another excellent tool to discriminate between true and false positives, only the wider p*I* range filtered datasets were used in the following.

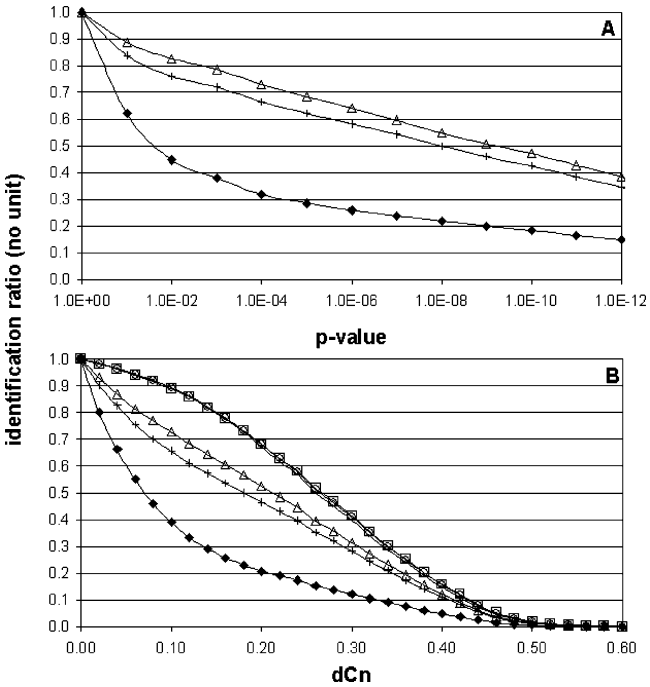

Figure 5. Ratio of identified peptides belonging to a scoring class. PHENYX results are displayed in panel A and SEQUEST results in panel B. Diamonds represent the untreated nr data, crosses the filtered data with ±2 p*I* stdev and open triangles with ±1 p*I* stdev tolerance, respectively. The *x*-axis range in panel A was limited to a p-value of 1E−12 for better illustration of differences. In panel B, the PeptideProphet peptides are displayed too, with open squares representing the untreated data, open circles and small filled circles the filtered data with ±2 and ±1 p*I* stdev tolerance, respectively.

Software Performance Comparison. Keller and colleagues recently compared the search results of three different search engines (SEQUEST, MASCOT, and COMET) and revealed that

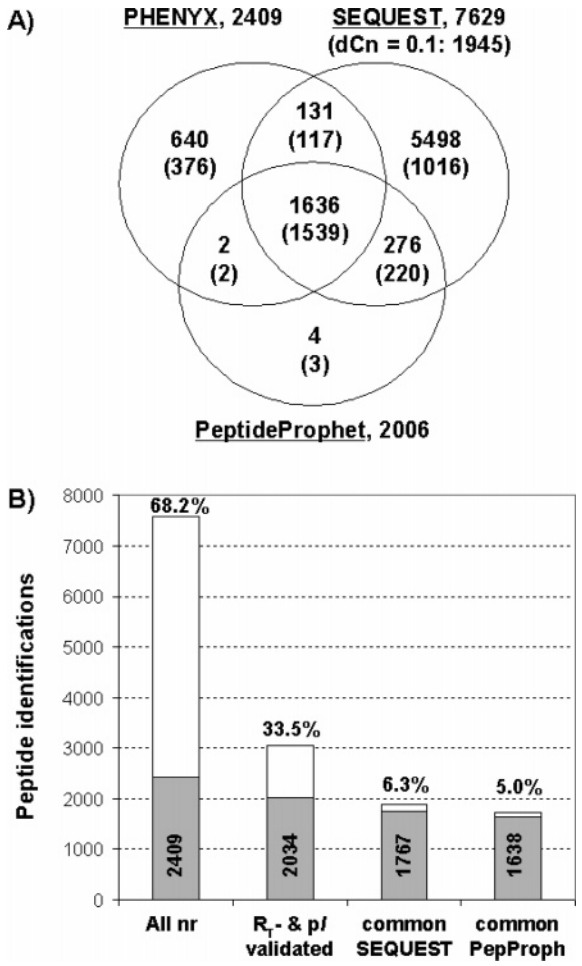

Figure 6. Comparison of peptide identification results from different search engines. The upper panel A is a comparison of all nr peptide identifications with the total number given behind the software name. Only PHENYX peptides with a *p*-value $< 1 \times 10^{-04}$ were considered. Numbers given in the diagram in parentheses correspond to identifications validated by the R_T- & 2 stdev p*I* filter. The bars in panel B represent the total number of nr PHENYX peptide identifications in the different datasets of the validation process. The lower light gray part corresponds to identifications with a *p*-value $< 1 \times 10^{-04}$ with absolute numbers given. The percentages given on top of each bar represent the relative amount of identified peptides with a *p*-value $\geq 1 \times 10^{-04}$ in each dataset.

close to 50% of the identifications were identical among the three.[31] The conclusion of their study was that 5% of incorrect results were common to two, and only 0.1% to all three, suggesting that the use of multiple search engines in combination can better distinguish between correct and incorrect search results with a possible loss of false negatives. It was therefore interesting to compare the PHENYX, SEQUEST, and PeptideProphet datasets in order to further validate the strength of our proposed strategy. The numbers of common identifications between the different search results are given in panel A of Figure 6. The two peptide identifications in the PeptideProphet set common with PHENYX but not with SEQUEST and the four identifications unique to PeptideProphet alone correspond to the six peptides that were removed from the SEQUEST list during cleaning from redundant entries as described above. Out of the 2409 PHENYX identifications with a *p*-value $< 1 \times 10^{-04}$, 84.4% were validated by the R_T and p*I* filter. Interestingly, this ratio was significantly increased to 93.7% in the subset of identifications common with SEQUEST results and it was down at 58.8% with the 640 identifications not confirmed by SEQUEST. These numbers clearly suggest that R_T and p*I* filtering is a powerful tool for the validation of peptide identifications. In absolute numbers, the R_T and p*I* validation process removes only 375 significant identifications with a *p*-value $< 1 \times 10^{-04}$ (15.6%) while many more false positives with a *p*-value $\geq 1 \times 10^{-04}$ were removed (4148 identifications = 80.2%) (panel B, Figure 6). Similar effects were achieved with SEQUEST identifications. In the subset of data common to SEQUEST and PHENYX the percentage of low scoring identifications was even further reduced (reduction rate of 97.7%) with the concomitant loss of well scoring peptides (26.7%), in other words false negatives. From this, it can be concluded that our proposed R_T and p*I* filtering criteria in combination with a PHENYX MS/MS spectra interpretation at a *p*-value cutoff of $< 1 \times 10^{-04}$ results in highly confident peptide identifications.

Additional Data Assessment. There were an additional 276 SEQUEST identifications validated by PeptideProphet that must be considered as true positives as well. Those 276 peptides were from 190 different database entries, of which 104 protein entries could not be found in the Uniprot-SProt/Tremble database searched with PHENYX. However, a manual check revealed that many protein sequences were actually present in Uniprot/Tremble but under a different accession number or with a different variant of high sequence identity. This was corroborated by the fact that 910 peptide identifications, corresponding to 149 proteins, were common on those "*non-Uniprot-Tremble*" proteins between PHENYX and SEQUEST results. The 41 proteins without any identification with PHENYX were based on a total of 184 peptide identifications in the nr SEQUEST results. Thereof, a combined number of 69 peptides were validated by PeptideProphet (52) or by the R_T- and p*I* filter (58), respectively. The average Xcorr and dCn scores for 1+, 2+, and 3+ ions were relatively high with 1.85 and 0.12, 2.36/0.17, and 2.96/0.23, respectively, suggesting that these identifications were mostly of good quality. The most likely explanation for such differences in search results is a different performance of the two computing algorithms. The following example illustrates this further. A potential asparagine-tRNA ligase protein (uniprot/tremble AC Q9V434, SEQUEST database AC AE003661) was identified and validated by SEQUEST and PHENYX with 7 peptides, albeit different ones. Three peptides were common to both datasets but neither search engine had a useful identification on those MS/MS spectra that were matched by the other one, although all spectra were of descent to good quality. Interestingly, PHENYX had no problem to correctly interpret three spectra previously not matched when the search was repeated against a reduced protein database containing Q9V434 together with all proteins identified in this study by PHENYX and validated by our R_T- and p*I*-filter criteria. This indicates that spectrum interpretation can fail when a large database is searched and combining the search results of two different search engines can rescue such losses due to different algorithms.

The different identification results are summarized in Table 2. The numbers of uniquely identified peptides and proteins were broken down according to the different data filtering criteria applied in this study. The first filter was based on a scoring criteria (*p*-value $< 1 \times 10^{-04}$ for PHENYX, and dCn $\geq$ 0.1 for SEQUEST results). While the *p*-value threshold strongly reduced the identifications in the raw, nonredundant data to

Table 2. Numbers of Unique Peptide and Protein Identifications before and after Data Treatment

	unique protein ID's						unique peptide ID's	
	Phenyx			Sequest				
	all	single ID's	in %	all	single ID's	in %	Phenyx	Sequest
no filter	1316	145	11.0	4254	1783	41.9	20420	33056
pVal $<1 \times 10^{-04}$	1024	495	48.3				2383	
dCn $\geq$0.1				1896	1021	53.9		4140
nonredundant data[a]	1205	411	34.1	1901	991	52.1	6874	7266
pVal $<1 \times 10^{-04}$	953	509	53.4				2095	
dCn $\geq$ 0.1[a]				945	439	46.5		2706
dCn $\geq$ 0.1[b]				499	183	36.7		1462
PeptideProphet validated				542	191	35.2		1717
PHENYX & SEQUEST[a]	755	382	50.6				1627	
and pVal $<1 \times 10^{-04}$	739	386	52.2				1534	
PHENYX & SEQUEST[b]	658	336	51.1				1316	
and pVal $<1 \times 10^{-04}$	652	348	53.4				1268	
PHENYX & PeptideProphet	720	366	50.8				1492	
and pVal <1E−04	711	376	52.9				1426	
R_T- and p*I*-filtered[a]	970	460	47.4	936	454	48.5	2736	2684
with pVal <1E−04	822	427	51.9				1772	
dCn $\geq$ 0.1[a]				608	261	42.9		1718
dCn $\geq$ 0.1[b]				484	190	39.3		1369
PeptideProphet				525	197	37.5		1601

[a] SEQUEST data filtering with relaxed criteria described in text, with at least partially tryptic peptides. [b] SEQUEST data filtering with more stringent criteria according to Qian et al., with at least partially tryptic peptides.[17]

a number close to numbers after more stringent filtering (2095 versus 1772 after R_T- and p*I*-filter, for instance) this is not true with the chosen dCn threshold. For this reason different groups proposed different recipes in order to filter false from true SEQUEST positives. We compared three variations, one with relaxed parameters (at least partially tryptic, Xcorr of 1.5, 2.0, 2.2 for +1, +2, and +3 ion charge state),[14] one with stringent criteria (Xcorr of 3.1, 3.8, 4.5 for +1, +2, and +3 ion charge state with partially tryptic peptides, and Xcorr of 1.5, 1.9, 2.9 for +1, +2, and +3 ion charge state with fully tryptic peptides),[17] and PeptideProphet validation, respectively. The latter two criteria removed significantly more identifications than the first at the risk of likely creating more false negatives, and the numbers of peptide identifications of 1462 and 1717 were similar to what was achieved with PHENYX and a more stringent p-value filter of $< 1 \times 10^{-07}$ (1580 peptides). The second filtering criteria was a trivial one and consisted in the removal of redundant identifications in case there were more than one potentially correct identification after above-mentioned decision trees. Such conflicts were solved by keeping the identification with the better p- or dCn value. The third filtering consisted of comparison of identification results between PHENYX and the other three SEQUEST datasets. As shown above, this is a good data validation criteria, but at the cost of neglecting many good identifications (only between 60.3 and 73.3% of nr PHENYX identifications with p-value $< 1 \times 10^{-04}$ overlapped with the different SEQUEST results). With the proposed fourth filtering strategy using experimentally derived R_T and p*I* criteria it was possible to accept absolutely more identifications of high confidence, but the different datasets behaved differently. The strongest reduction of 36.5% was observed with the relaxed SEQUEST results (from 2706 entries with dCn $\geq$ 0.1 down to 1718). PHENYX was reduced by 15.4% (2095 with p-value $< 1 \times 10^{-04}$ down to 1772), and stringent SEQUEST and PeptideProphet by a little more than 6% each (1462 down to 1369, and 1717 to 1601, respectively). It should be noted that PeptideProphet validated results were dCn independent. The inverse of these reduction rates (1 minus reduction rate) can be interpreted as the confidence level of each dataset with the relaxed SEQUEST data performing the poorest. In terms of protein identifications 945 proteins were identified with this set without the p*I* and R_T filter applied. After filtering, this number was reduced to 608 proteins, hence the original result contained something like 35.6% of potentially wrong protein identifications. PHENYX performed significantly better with about 13.7% of wrong protein identifications using a very relaxed p-value threshold of $< 1 \times 10^{-04}$. Regarding the numbers of identified proteins, it is also interesting to mention that PHENYX identified more proteins (822 identifications) with less peptides (2.2 peptides per protein in average) than any of the SEQUEST strategies with 608 (2.8 peptides/protein) in case of relaxed, 484 (2.8 peptides/protein) with stringent parameters, and 525 (3 peptides/protein) with PeptideProphet.

The numbers in Table 2 belong to unique, single peptide and protein identifications. For instance, in the case of PHENYX, we have claimed over 2034 peptide identifications after R_T- and p*I*-validation. This number was reduced to 1772 unique peptides as 13.3% of the peptides were identified in two OGE fractions, and 1.4% of the peptides were identified in three to six OGE fractions. The latter population of badly focusing peptides could be divided into two categories. The first contained 13 peptides having a neutral and badly defined p*I* due to a charge to pH titration curve spanning several pH units at zero charge. These peptides were identified in the OGE fractions 4 to 11. The second contained 2 acidic and 9 basic peptides identified in OGE fractions 1 to 3 and 12 to 15, respectively. Interestingly, both groups of badly focused peptides were composed of relatively more proline residues than two randomly chosen, representative sets of well focused peptides with a similar number of amino acid residues (Figure 7 in Supporting Information). Proline is known to introduce bends in the two-dimensional structure of peptides. Such bends may enforce 3-dimensional conformations that influence the ionic strength of charged groups, changing pK_a values and consequently the isoelectric point of the peptide.

Discussion

We introduce here a new strategy for shotgun proteomics projects that allows for completely orthogonal peptide fractionation in two dimensions, namely isoelectric focalization followed by reverse phase LC−MS/MS combined with improved peptide identification validation based on experimentally derived peptide parameters. We used OGE for the first dimension that has the inherent advantage of recovering peptides directly in solution without contamination by oil or ampholytes in contrast to the recently advocated use of IPG strips where several washing and extraction steps were needed to extract the peptides from the gel matrix.[21] The experimentally derived parameters of isoelectric point and hydrophobicity were used to validate the identification results. The concept of validating peptide identifications by experimental data has already some history. The most straightforward approach to validate the correctness of a peptide identification is to use the accurate mass measured by the mass spectrometer. This feature is already part of any database search engines such as SEQUEST, MASCOT, and most recently OLAV/PHENYX[18−20] because the rate of false identifications can be reduced enormously by restricting the search to a few ppm mass error. Unfortunately, measuring exact peptide masses with only a few ppm mass error affords expensive instruments, such as FT−ICR MS. For this reason, iontrap mass spectrometers are still the preferred work horses with the inherent disadvantage of producing less accurate mass data. Another concept already long in use and perfected by many is the comparison of experimentally measured and theoretically calculated RPC retention time of peptides.[24,29] More recently, it was also realized that peptide p*I* values could also serve the purpose of validating peptide identifications.[13,14,21] Each of these parameters alone can already improve the validation process. However, the accurate theoretical prediction of p*I* or R_T is difficult because they depend on many experimental parameters coming from differences of the surrounding matrix, like ionic strength of solution, denaturing agents, ampholytes, RPC solvent composition, impurities, stationary phase, or chemical and physical properties of the peptides themselves, like deamidation, oxidation, rotational freedom of peptide bonds, peptide folding energies, intermolecular peptide affinities etc. to name a few. Hence, there is always a need to accept a certain error tolerance in order not to discard true positives. This variability between theoretically predicted and experimentally measured parameter is nicely reflected in Figures 2 and 3. By applying more than one acceptance criteria, the likelihood that falsely identified peptides pass all the filters becomes smaller and the choice of more relaxed acceptance criteria boundaries can be accepted in order to allow for the experimental variability. This assumption could be verified with our approach where we compared the identification results after R_T- and p*I*-filter validation with a validation based on identification comparison between SEQUEST/PeptideProphet and PHENYX. Out of experience by manually checking identification results and from the results displayed in Figure 4, it was concluded that a PHENYX *p*-value threshold of smaller than 1×10^{-04} provides peptide identifications with a probability of about 95% to be correct. PHENYX returned 2409 nr peptide identifications (nr in each OGE fraction subset) in this study that were fulfilling this *p*-value criteria. Only about 11% of these PHENYX identifications could not be validated by any approach and 68.8% were confirmed by both our relaxed R_T- and p*I*- filter criteria and SEQUEST (Figure 6). The roughly 20% of peptides that were validated by only one approach reflect very likely differences caused by searching different sequence databases as shown in the Results section. The PHENYX *p*-value is a powerful tool to reduce the level of false positives, but most likely to create more false negatives too. This is demonstrated by changing the *p*-value threshold to a more stringent threshold of $< 1 \times 10^{-07}$. With that threshold in place only 2.7% of identifications are not validated by any method and a comparatively bigger ratio of 80.8% was validated by the R_T-p*I*-filter and SEQUEST but at the cost of a reduced proteome coverage, with 608 peptides lost. The proportion of R_T- and p*I*-validated peptides did increase less significantly from 84.4% to 92.2% (results not shown). From this, it can be concluded that the ratio of peptide identifications validated by our R_T-p*I*- filter in each data set is likely a direct measure for the confidence in having correct true positive results. This ratio was with 94% (Table 2) highest with PeptideProphet and the stringent filter criteria for SEQUEST results, corroborating the observation made in Figure 1, panel C. PHENYX with a *p*-value cutoff of 1×10^{-04} performed less well with 84.4% validated identifications, while there was a rather high percentage of apparently false results in the SEQUEST set, which was filtered with relaxed scoring parameters, where only 65% of identifications passed the filter criteria. Our results, together with the statistical evaluation on SEQUEST performed by the group of R. Smith,[17] indicate that many of the recently published results on proteome characterizations by a shotgun approach and SEQUEST as peptide identification software have to be treated carefully.

Although our data validation criteria for isoelectric point and RPC retention time were set relaxed we could show with the different data validations applied that a filter based on a combination of R_T and p*I* values is a powerful tool for the automatic validation of peptide identifications. From our work, as well as that from the work of others, it also becomes obvious that a peptide identification software producing a score with a relation to the statistical probability about the correctness of the results, like the *p*-value in PHENYX, is of high value. Peptide identification with PHENYX in combination with our proposed experimental data validation approach is a powerful tool that should enable to work with proteomes where only badly annotated genomes are available. The R_T- and p*I*-based validation process would become even more powerful if the R_T and p*I* calculation machine would be trained on a neuronal network using well validated data acquired on a standardized analytical system, e.g., as was done for the normalized elution time values (NET) by the group of R. Smith.[24]

Acknowledgment. The authors thank Agilent Technologies Deutschland GmbH for financial support and fruitful collaboration on Off-Gel electrophoresis. We also thank James Eddes and Jimmy Eng of the Institute of Systems Biology (Seattle, WA) and Alexandre Masselot of GeneBio (Geneva, Switzerland) for their assistance with SEQUEST or PHENYX searches.

Supporting Information Available: The redundant data, a total of 7582 and 7629 peptides for PHENYX and SEQUEST, respectively (Supporting Information Tables 1, 2, and 3). Representative sets of well focused peptides with a similar number of amino acid residues (Supporting Information Figure 7). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Wilkins, M. R.; Sanchez, J. C.; Gooley, A. A.; Appel, R. D.; Humphery-Smith, I.; Hochstrasser, D. F.; Williams, K. L. *Biotechnol. Gene Eng. Rev.* **1995**, *13*, 19−50.

(2) Gygi, S. P.; Corthals, G. L.; Zhang, Y.; Rochon, Y.; Aebersold, R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9390−9395.

(3) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R. III. *Nat. Biotechnol.* **1999**, *17*, 676−682.

(4) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994−999.

(5) Gygi, S. P.; Rist, B.; Griffin, T. J.; Eng, J.; Aebersold, R. *J. Proteome Res.* **2002**, *1*, 47−54.

(6) Rose, K.; Bougueleret, L.; Baussant, T.; Böhm, G.; Botti, P.; Colinge, J.; Cusin, I.; Gaertner, H.; Gleizes, A.; Heller, M.; Jimenez, S.; Johnson, A.; Kussmann, M.; Menin, L.; Menzel, C.; Ranno, F.; Rodriguez-Tomé, P.; Rogers, J.; Saudrais, C.; Villain, M.; Wetmore, D.; Bairoch, A.; Hochstrasser, D. *Proteomics* **2004**, *4*, 2125−2150.

(7) Pieper, R.; Gatlin, C. L.; Makusky, A. J.; Russo, P. S.; Schatz, C. R.; Miller, S. S.; Su, Q.; Mcgrath, A. M.; Estock, M. A.; Parmar, P. P.; Zhao, M.; Huang, S. T.; Zhou, J.; Wang, F.; Esquer-Blasco, R.; Anderson, N. L.; Taylor, J.; Steiner, S., *Proteomics* **2003,** *3*, 1345−1364.

(8) Wall, D. B.; Kachman, M. T.; Gong, S.; Hinderer, R.; Parus, S.; Misek, D. E.; Hanash, S. M.; Lubman, D. M. *Anal. Chem.* **2000**, *72*, 1099−1111.

(9) Fung, K. Y. C.; Askovic, S.; Basile, F.; Duncan, M. W. *Proteomics* **2004**, *4*, 3121−3127.

(10) Strader, M. B.; Verberkmoes, N. C.; Tabb, D. L.; Connelly, H. M.; Barton, J. W.; Bruce, B. D.; Pelletier, D. A.; Davison, B. H.; Hettich, R. L.; Larimer, F. W.; Hurst, G. B. *J. Proteome Res.* **2004**, *3*, 965−978.

(11) Heller, M.; Stalder, D.; Schlappritzi, E.; Hayn, G.; Matter, U.; Haeberli A. *Proteomics* **2005**, *5*, 2619−2630.

(12) Michel, P. E.; Reymond, F.; Arnaud, I. L.; Josserand, J.; Girault, H. H.; Rossier J. S. *Electrophoresis* **2003**, *24*, 3−11.

(13) Heller, M.; Michel, P. E.; Crettaz, D.; Wenz, C.; Tissot, J. D.; Reymond, F.; Rossier, J. S. *Electrophoresis* **2005**, *26*, 1174−1188.

(14) Cargile, B. J.; Bundy, J. L.; Stephenson, J. L. Jr. *J. Proteome Res.* **2004**, *3*, 1082−1085.

(15) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.

(16) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. *J. Proteome Res.* **2003**, *2*, 137−146.

(17) Qian, W. J.; Liu, T.; Monroe, M. E.; Strittmatter, E. F.; Jacobs, J. M.; Kangas, L. J.; Petritis, K.; Camp, D. G.; Smith, R. D. The Human Proteome. *J. Proteome Res.* **2005**, *4*, 53−62.

(18) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. *Proteomics* **2003**, *3*, 1454−1463.

(19) Eng, J. K.; McCormack, A. L.; Yates, J. R. III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(20) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.

(21) Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L. Jr. *J. Proteome Res.* **2004**, *3*, 112−119.

(22) Martin, A. J. P. *Ann. NY Acad. Sci.*, **1948**, *49*, 249−264.

(23) Sakamoto, Y.; Kawakami, N.; Sasagawa, T. *J. Chromatogr.* **1988**, *442*, 69−79.

(24) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. *Anal. Chem.* **2003,** *75*, 1039−1048.

(25) Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G. 2nd; Smith, R. D. *J. Proteome Res.* **2004**, *3*, 760−769.

(26) Echalier, G. *Drosophila Cells in Culture***;** Academic Press: Toronto 1997.

(27) Yi, E. C.; Lee, H.; Aebersold, R.; Goodlett, D. R. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2093−2098.

(28) Bjellqvist, B.; Hughes, G. J.; Pasquali, C.; Paquet, N.; Ravier, F.; Sanchez, J. C.; Frutiger, S.; Hochstrasser, D. *Electrophoresis* **1993**, *14*, 1023−1031.

(29) Guo, D.; Mant, C. T.; Taneja, A. K.; Parker, J. M. R.; Hodges, R. S. *J. Chromatogr.* **1986**, *359*, 499−517.

(30) Lehninger, A. *Biochemistry*, 3rd ed.; Worth Publishers: New York, 1985.

(31) Keller, A.; Eng, J.; Zhang, N.; Aebersold, R. *Mol. Systems Biol.*, in press.

PR050193V