Published in final edited form as:

Anal Chem. 2011 July 1; 83(13): 5259–5267. doi:10.1021/ac200609a.

Preview: A Program for Surveying Shotgun Proteomics Tandem Mass-Spectrometry Data

Yong J. Kil^{1,4}, Christopher Becker⁴, Wendy Sandoval², David Goldberg^{1,3}, and Marshall Bern^{1,4}

¹ Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304

² Department of Protein Chemistry, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080

⁴ Protein Metrics Inc., 3333 Coyote Hill Rd., Palo Alto, CA 94304

Abstract

Database-search programs for peptide identification by tandem mass spectrometry ask their users to set various parameters, including precursor and fragment mass tolerances, digestion specificity, and allowed types of modifications. Even proteomics experts with detailed knowledge of their samples may find it difficult to make these choices without significant investigation, and poor choices can lead to missed identifications and misleading results. Here we describe a program called Preview that analyzes a set of mass spectra for mass errors, digestion specificity, and known and unknown modifications, thereby facilitating parameter selection. Moreover, Preview optionally recalibrates mass-over-charge measurements, leading to further improvement in identification results. In a study of Bruton's tyrosine kinase, we find that the use of Preview improved the number of confidently identified mass spectra and phosphorylation sites by about 50%.

Introduction

Shotgun or "bottom-up" proteomics analyzes complex protein mixtures by digesting proteins with a protease such as trypsin and then identifying the resultant peptides using tandem mass spectrometry (MS/MS). There are a number of computational search tools to support peptide identification by MS/MS; the three most widely used are Mascot¹, SEQUEST², and X!Tandem³. These programs compare observed fragmentation spectra to predicted fragmentation spectra for peptides from a database of protein sequences. The user of a search program must input various parameters:

Mass tolerances

The user sets tolerances for precursor and fragment masses that reflect the type of MS/MS instrument and the data acquisition strategy. For example, with Orbitrap⁴ MS and linear ion-trap MS/MS, the user may configure the program to consider peptides with mass within 10 ppm of the measured precursor mass and to score fragment ions with mass over charge (m/z) within 0.4 Daltons per charge of the measured m/z of a peak.

Address Correspondence to: Marshall Bern, Ph.D., Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304, bern@parc.com, Voice: (650) 812-4443, Fax: (650) 812-4471.

³Current address: eBay Inc., 2065 Hamilton Ave., San Jose, CA 95125

Digestion specificity

The user may set the program to consider only peptides with digestion-specific cleavages at both termini (for trypsin, after arginine and lysine), or the user may choose a broader search, allowing missed cleavages and nonspecific cleavage at one or both termini.

Modifications

The most difficult choice for the user is which peptide modifications⁵ to allow. Some *in vitro* modifications are ubiquitous, occurring to some extent in almost all shotgun proteomics samples, but others depend upon the sample and preparation and can vary unpredictably. Posttranslational modifications (PTMs) also vary from sample to sample and from protein to protein within a sample. If the user searches for more than about 8 variable modifications, meaning modifications that may or not be present at each site, the search may be impractically slow and give many identifications that are either partially or completely false as the search size explodes⁶.

Some existing search engines already offer partial solutions to the problem of parameter setting. Mascot's error-tolerant search⁷ considers nonspecific cleavage along with a large number of modifications, but limits the search to modifications included in Unimod⁵ and allows only one "anomaly" per peptide. Paragon⁸ allows multiple modifications per peptide, but only on the most promising peptides. InsPecT⁹ offers "blind" modification search,^{10–12} which allows arbitrary mass shifts, but blind search tends to be slower and less accurate than known modification search, because it does not take advantage of protein chemistry knowledge. MODi includes both known and blind modification search,¹³ but can handle only a limited number of proteins. Spectrum-to-spectrum comparison, as in Modificomb¹⁴, Bonanza¹⁵, or spectral networks analysis¹⁶, improves the speed and accuracy of blind modification search, but can only identify modified peptides that are also observed without modifications.

Here we describe a new tool called Preview that offers a more complete solution. Preview has only two required inputs: a set of MS/MS spectra (in .mgf or .dta formats) and a protein database (in FASTA format). As shown in Figure 1, the program measures precursor and fragment m/z errors, estimates the amount and type of nonspecific digestion, assays the prevalence of known modifications, and reports unrecognized (blind-search) modifications. The user can then set the parameters for a conventional search engine based upon Preview's statistics and the aims of the proteomics project. Preview optionally recalibrates m/z measurements and outputs a new .mgf or .dta file.

Preview operates in a fraction of the time of a standard search program; for example, a complete search of the Aurum¹⁷ data set (9987 MS/MS spectra) against a database containing ~90,000 protein sequences took 93 seconds, less than one-fiftieth the time (92 minutes) of an eight-modification search using X!Tandem, the fastest of the commonly used search programs. In order to achieve this speed, we made a number of simplifying assumptions in the design of Preview. The foremost assumption is that the 100 most detectable proteins faithfully represent the entire sample for the full menu of search parameters. This holds true for simple samples, containing less than 100 proteins. For complex samples, the assumption should be accurate for estimations of m/z errors, nonspecific digestion, and *in vitro* modifications, but less accurate for *in vivo* modifications, which are usually protein-specific. Another simplification is that Preview, with some exceptions, searches for unrelated types of modifications one at a time, thereby avoiding the combinatorial explosion of multiple modification searches. Finally, Preview's peptide identification algorithm takes shortcuts: like SEQUEST, it represents both predicted and observed peaks by integer masses, so that scoring a candidate peptide against a spectrum can

be done with two instructions per predicted peak, regardless of the number of observed peaks.

Despite the loss of sensitivity incurred by these simplifications, we have found Preview valuable in our own bioinformatics pipeline, which uses Byonic¹⁸ as the primary search program. Preview can be used in conjunction with any search program; one X!Tandem user reports 10% better protein sensitivity on TOF-TOF data due to Preview's m/z recalibration. Preview can also improve laboratory practices and reproducibility by providing timely feedback on instrument calibration and sample preparation artifacts. Preview is now available for public use via free download from www.proteinmetrics.com. There are versions for Windows, Mac, and Linux / Unix, for both 32- and 64-bit architectures.

Methods

Peptide/Protein Identification

Preview contains many of the standard steps of shotgun proteomics bioinformatics: peptide scoring (as in Mascot), protein assembly (as in ProteinProphet¹⁹), and blind modification search (as in Popitam¹⁰ or InsPecT⁹). We compared Preview with Byonic¹⁸ (in-house software) and X!Tandem (version TORNADO 2010.01.01.2). Unless otherwise noted, we used the IPI.Human.6102008.fa protein database. We allowed semitryptic cleavage with one missed cleavage for X!Tandem and any number of missed cleavages for Byonic. X!Tandem does not allow more than one variable ("potential") modification per residue type, so X!Tandem had to be run several times to search for M[+16], M[+32], W[+16], and W[+32].

Spectrum Preprocessing

Preview takes as input a set of centroided MS/MS spectra, either in Mascot generic format (MGF) or as .dta peak lists, and a protein database in FASTA format. The first step of the program converts the spectra into a simplified form, in which each spectrum is represented by a floating-point precursor m/z value, an integer charge, an integer precursor mass, 2000 integer-valued m/z bins, and 8 floating-point m/z's of intense fragment peaks. The scoring step uses only the precursor value and the m/z bins; the 8 floating-point masses are included to estimate m/z measurement errors. If the user indicates that precursor charges are uncertain, as is often the case with ion-trap spectra, then Preview will run each spectrum assuming charges +1, +2, and +3.

To compute the simplified spectrum, Preview first extracts the 300 most intense peaks in the spectrum, and then downweights isotope peaks as previously described for Byonic¹⁸. Observed floating-point m/z's are then converted to integer m/z bins by rounding to remove mass defects (the fractional parts of elemental masses). A value of M is rounded to the closest integer to 0.9995 M, so that an observed value of 1814.1 rounds to 1813. This method of rounding takes advantage of the characteristic mass defects of peptide masses, so that multiple observations of the same fragment ion almost always round to the same integer bin.

We denote an experimental spectrum by X(i), where i is the m/z bin, running from 1 to 2000. The X(i) values, which we call "weights", are not proportional to peak intensities in the original spectrum, but rather they reflect intensity relative to peaks of similar m/z. Each m/z bin is represented by a signed byte (8-bit integer), but we currently only use integers in the range from -3 to 6. A bin i without a peak is given a negative X(i) in order to penalize candidate peptides for predicting theoretical peaks that are not observed. Details are given in the Supplemental Information.

Peptide Scoring

Preview keeps all the spectra in memory and then scores each candidate peptide against all the spectra of the right mass. Preview uses a dot-product score: $\sum_i T(i) * X(i)$, where $T(i)$ is the theoretical weight (an integer value from 0 to 3) and $X(i)$ is the experimental weight at mass i (an integer value from -3 to 6).

For collision-induced dissociation (CID), the theoretical weight $T(i)$ is set to 1 for b-ions, starting at b_2 and running up to 2000 Daltons. Preview also scores b-ions with single water losses if the ion contains serine or threonine; these ions also have theoretical weight 1. Preview gives the most commonly observed a-ions, a_2 and a_4 , weight 1, but does not score any other a-ions. Preview gives theoretical weight 1 to all y-ions starting at y_1 , except for y-ions with mass between 300 and 1000 Da, which have weight 2. Except for neutral losses from the precursor, Preview does not score multiply charged ions. Doubly charged y-ions are common in CID spectra of multiply charged precursors, but we chose not to score them because rounding them to integers gives unpredictable results and empirically worse performance; for example, 1000.0, 1000.5, and 1001.0 are all likely m/z 's for a doubly charged ion.

To compute the dot-product, we do not need to perform 2000 multiplies and adds, because there are typically fewer than 50 nonzero $T(i)$'s. We compute a list of i values with nonzero $T(i)$, so that i appears on the list once in the case $T(i) = 1$ and twice in the case $T(i) = 2$. We then use this list to index into the $X(i)$ array, so that the dot-product can be computed with about 50 indexing operations (table lookups) and 50 integer additions. For still greater speed, we “prescore” a candidate peptide using just 10 or 11 of the most commonly observed theoretical ions, each with weight one: b_2 to b_6 and y_3 to y_8 for CID and c_4 to c_8 and z_4 to z_8 for ETD (electron transfer dissociation). If a candidate does not achieve a sufficient score (usually set to -4) on these 10 or 11 ions, then the candidate is rejected without computing the full score. Less than 2% of the candidates go on to full scoring, so scoring a typical candidate takes about 10 indexing operations, 10 integer additions, and one integer comparison.

Preview uses two score thresholds: $T_{\text{High}} = \max \{ s+1, 23 \}$, where s is the maximum score achieved by any unmodified decoy peptide in the initial search; and $T_{\text{Low}} = \max \{ t+1, 15 \}$, where t is the maximum initial-search score achieved by any unmodified decoy peptide with at least 9 residues. The higher score threshold is used for searches over a large number of peptides, such as a search for N-terminal acetylation, and the lower threshold is used for smaller searches, such as a search for N-terminal pyro-glu from glutamine, which only applies to peptides beginning with Q; the exact constants of 23 and 15 were chosen empirically for good performance. No matter which threshold is used, Preview corrects the number of hits to target peptides by subtracting off the number of hits to decoy peptides, so that the corrected number estimates the expected number of true target peptide hits.

Protein Ranking and Peptide Database Assembly

Preview performs an initial search to compute a list of representative proteins. By default, the initial search is a fully tryptic search, but the cleavage residues and the specificity requirements can be set by the user. Preview scores proteins twice, first to assign ambiguous spectra—those matching peptides that appear in more than one protein—to unique proteins, and second to assemble the list of representative proteins.

The first score for a protein is simply the sum of all the spectrum scores for spectra matching peptides in that protein. Thus in the first score, an ambiguous spectrum contributes to more than one protein. If two proteins end up with exactly tied scores, the tie is broken by protein number, meaning the order that the proteins appear in the initial protein database. After

computing the first score, Preview assigns each ambiguous spectrum to the highest-scoring protein containing the matched peptide. The second score for protein A is the sum of four terms: twice the number of distinct peptides with scores over 60, the number of distinct peptides with scores between 30 and 59, a small constant (0.02) times the total score of all spectra matching that protein, and a negative correction for the length of the protein, $-\text{Frac(A)} \cdot \text{TotalScore}$, where Frac(A) is the fraction of the protein database occupied by protein A and TotalScore is the total score of all spectra matched. The dominant factor in this sum is usually the second one. The constants of 60, 30, and 0.02 were chosen empirically for good performance over a variety of data sets.

Preview ranks proteins by descending order of second score and then cuts off the list when the protein score falls below a threshold, the number of proteins reaches 100, or the number of deliberate decoys (recognized by protein names in the FASTA database that start with the string >Reverse) reaches two. The proteins remaining on the list are the **representative proteins**.

After choosing the representative proteins, Preview assembles a peptide database of up to 22,000 peptides. The peptide database initially contains all digestion-specific peptides from the top-ranking representative proteins, regardless of the number of missed cleavages. Because the initial protein residue is often removed *in vivo*, Preview considers this cleavage as digestion-specific for all types of digestion. Assay A.3 below augments the peptide with semi- and non-specific peptides. The peptide database also includes a matched set of decoys, for use in the target/decoy approach to false discovery rate (FDR) estimation.²⁰ Whenever Preview adds a peptide to the peptide database, it also adds the peptide with all its residues except the last one in reverse order; that is, the “reverse” of the peptide TIFIISMYK is YMSIIFITK.

Computational Assays

Preview performs a number of searches to test for mass accuracy, cysteine treatment, digestion specificity, and modifications. The searches are ordered so that the ones with results most likely to affect subsequent searches are performed first. For example, Preview checks mass accuracy first in order to set tolerances for subsequent searches. Preview checks the cysteine treatment next, including artifacts such as over-alkylation²¹. For simplicity, later searches do not affect earlier searches, so that if Preview discovers, for example, widespread over-alkylation it does not go back and re-estimate mass accuracy. Throughout its searches Preview stores, for each spectrum, the best score achieved so far. An identification must achieve a score at least as good to be considered valid. Thus Preview rejects an identification of MP[+16]EPTYK when assaying for hydroxyproline if M[+16]PEPTYK, previously considered in the assay for oxidized methionine, scored higher, but will accept MP[+16]EPTYK if the two integer-valued scores are tied. This design choice mitigates but does not eliminate the dependence of the results on search order.

In each of its assays, Preview considers only the peptides in the database relevant to the assay. A preliminary pass through the peptide database marks each peptide as valid or invalid for the upcoming search. For example, in assaying for oxidized methionine, Preview considers only methionine-containing peptides. This restriction gives a modest speed improvement; it also enables more uniform output, with percentages taken relative to the number of peptides on which the modification could occur. In these restricted searches, spectra from currently invalid peptides or from peptides not included among the representative proteins, will give low scores, and will presumably match target and decoy peptides at the same rate, so that they will not bias the statistics.

Assay results are presented on two HTML pages, a Summary page and a Detailed Results page. The Summary page shows the top 10 proteins, plots and statistics of m/z measurement errors and recalibration, digestion specificity, fixed modifications, and the most common variable modifications. The Detailed Results page shows all the representative proteins, scoring statistics, and organizes the assay results into seven sections, described in A1 – A7 below.

A1. Mass Accuracy

Preview computes m/z errors from the theoretical and the observed m/z values. The program reports precursor m/z errors (median absolute error, median relative error in parts per million, etc.) over all digestion-specific, unmodified peptide identifications with score at least T_{High} . For fragment m/z errors, Preview checks the 8 intense real-valued peaks in the simplified form of the spectrum, and uses a peak with mass m for error statistics if its integer part (closest integer to 0.9995 m) matches the integer part of a theoretical singly-charged b- or y-ion. If errors exceed 0.5 Daltons, Preview may match the $^{13}\text{C}_1$ isotope peak to the theoretical monoisotopic peak, but this problem is rare due to the spectrum preprocessing, which downweights isotope peaks.

In order to recalibrate m/z measurements, Preview writes out two temporary files, one for precursor ions and one for fragment ions, containing pairs of numbers, (Observed m/z, Observed m/z – Theoretical m/z). Preview then uses two rounds of least-squares regression to fit a quadratic polynomial to the pairs. Pairs with residual error in the worst quartile are removed after the first round, in order to guard against outliers caused by incorrect identifications. The final quadratic correction curves are used to produce a new spectrum file in either .mgf or .dta format. Figure 2 shows the plotted pairs, including some outliers in the precursor m/z errors, along with the final correction curves.

A2. Cysteine

If the user has not specified a fixed modification for cysteine, Preview next performs a series of searches (with threshold T_{Low}) to determine the cysteine treatment. It searches the following fixed modifications: C[+0] (no treatment), C[+46] (beta-methylthiolated, MMTS treatment), C[+57] (carbamidomethylated, iodoacetamide treatment), C[+58] (carboxymethylated, iodoacetic acid treatment), and C[+71] (propionamide). If at least four cysteine-containing peptides are found with one of these options, then Preview declares the option with the most identifications to be the fixed cysteine modification. No matter which fixed modification, Preview looks for unmodified cysteine, denoted C[+0], unbroken disulfide bridges (which appear as C[−2] in peptides containing at least two cysteines), and cysteine propionamide, C[+71], a common artifact of acrylamide gel electrophoresis. For iodoacetamide treatment, Preview considers N-terminal +57 and +114, H[+57], and K[+57], and reports the percent of peptide identifications carrying at least one of these over-alkylation artifacts. For iodoacetic acid treatment, Preview checks for N-terminal +58 and M[+44].

For all its assays, Preview reports both a percentage and a fraction, for example, “C[+71] 13.0% (7/54)” means that 7 of 54 identifications of cysteine-containing peptides contain at least one C[+71] modification. The denominator of the fraction is the number of identified peptides that could possibly contain the modification, in this case, cysteine-containing peptides. The fraction (along with the total number of spectra) implies statistical sampling error, so that the user knows to trust “C[+71] 12.5% (1/8)” less than “C[+71] 13.0% (7/54)”. Duplicates count, so that 7 peptides with C[+71] could be 7 different peptides or 7 instances of the same peptide. This policy is consistent with Preview’s model of modifications as random events, the same for all peptides.

A3. Digestion Specificity

Preview next estimates digestion specificity. We explain the process assuming a trypsin digestion, but the program allows more general cleavage rules. Preview searches the spectra against “N-ragged” (nonspecific at the N-terminus) and “C-ragged” semitryptic peptides from the representative proteins, counts the number of hits with score above T_{High} , and then corrects for the number of false hits estimated by the target/decoy approach. Preview adds each semitryptic peptide with score above 15, along with its reverse peptide, to the peptide database. Preview then does the analogous search for fully nontryptic peptides, and augments the database with nontryptic peptides scoring at least 15, along with their reverses.

A4. Oxidation

Preview divides non-cysteine peptide modifications into three logical types: oxidations, “chemical modifications”, and biological PTMs. Oxidations are variable modifications, but they tend to follow a pattern, so that lightly oxidized samples typically have only methionine sulfoxide, M[+16], and more heavily oxidized samples have M[+16], H[+16], W[+16], M[+32], W[+32], and C[+48], and possibly many other forms. Preview takes advantage of this pattern by assaying some co-occurring oxidations together. For all modifications, Preview reports the percentage of peptides with that modification relative to the number than could have that modification, for example, for oxidized methionine, Preview reports $\max \{ 0, n_{\text{M16}} / n_{\text{M}} \}$, where n_{M16} is the number of peptide identifications containing at least one M[+16] and the denominator n_{M} is the total number of identifications containing at least one M, oxidized or not, with score exceeding T_{Low} .

A5. Chemical Modifications

Preview allows the user to specify fixed modifications (chemical derivitizations) on cysteine, lysine, arginine, the peptide N-terminus, or the peptide C-terminus. Preview assays each fixed modification for completeness by searching with and without the modification. For example, in assaying for fixed lysine acetylation Preview restricts the search to lysine-containing peptides, and reports $n_{\text{K42}} / (n_{\text{K}} + 0.0001)$ as the “completeness”, where n_{K42} is the number of identifications with score above T_{Low} in a search assuming fixed K[+42] and n_{K} is the number of identifications containing either K[+42] or unmodified K.

Preview also reports the following as “chemical modifications”: deamidation of N and Q; amidation of D and E; pyro-glu from N-terminal Q, E, and carbamidomethylated cysteine (camC); sodiation of any one residue; carbamylated R and M; lysine, histidine and N-terminal +26 from acetaldehyde; N-terminal methylation and dimethylation; N-terminal acetylation; E[+14] (methyl esterification); and formylation of S and T. These modifications are all assayed with searches and statistics analogous to the ones described above for oxidized methionine.

A6. Posttranslational Modifications

Preview assays the following PTMs: hydroxyproline; phosphorylation of S, T, and Y; beta-elimination of S and T (not actually a biological PTM, but possibly a marker of phosphorylation or O-linked glycosylation); methylation of K, H, N, and R; dimethylation of K and R; and acetylation of K and protein N-terminus.

A7. Unanticipated Modifications

Finally Preview searches for unanticipated modifications using a “wild card” modification as in Byonic¹⁸ or Protein Prospector²². A wild card allows any integer mass within a range, to be added to any one residue in a peptide. By default, Preview uses a mass range of -50 to +150 Da. A wild-card modification of mass m shifts all the peaks for ions containing the

modification by m, that is, T(i+m) receives the value of T(i) in the theoretical spectrum of the unmodified peptide. Preview retains only high-scoring wild-card identifications, and reports the identified sequences to the user in a separate spreadsheet for expert inspection.

Sample Preparations

We compared Preview's results in detail to those from two conventional search programs, Byonic and X!Tandem, on three well-characterized MS/MS training data sets, the first three data sets listed in Table 1. Then we applied Preview to a study of Bruton's tyrosine kinase, and greatly improved the depth of data analysis compared to initial analyses at Genentech and PARC. All data sets are publicly available on Tranche / Proteome Commons

Benchmark Data Sets

The first three data sets were collected as training data by three different laboratories, and represent three different sample preparation methodologies and MS/MS instruments. As described elsewhere¹⁷ the Aurum sample consists of 246 recombinantly manufactured human proteins, individually purified using both affinity tags and SDS-PAGE, alkylated with iodoacetamide, and individually digested with trypsin. The spectra were acquired on an ABI 4700 MALDI TOF/TOF instrument.

The Jurkat data set is from Genentech (South San Francisco, CA). Jurkat cells were lysed using 8M urea/50 mM Tris (pH 7.5) in the presence of protease and phosphatase inhibitors (Roche), then separated by SDS-PAGE. Proteins of mass at least 70 kDa were digested in gel with trypsin, extracted, and dried. Approximately 1 µg of digested peptides was injected for analysis on a Thermo LTQ Orbitrap with Orbitrap single-MS scans and LTQ MS/MS scans. This sample had no cysteine treatment.

We obtained the Plasma data set from PPD, Inc. (Menlo Park, CA), now part of Caprion Proteomics. The sample is human blood plasma, depleted of 6 abundant proteins using a multiple-affinity removal system (Agilent), reduced, alkylated with iodoacetic acid, trypsin-digested, and analyzed by capillary LC-ESI-MS/MS (with 0.1% formic acid in the solvent) on a Thermo LTQ instrument. Cysteines are carboxymethylated.

Bruton's Tyrosine Kinase

We applied Preview to research on Bruton's tyrosine kinase, a well-studied enzyme²³ implicated in X-linked agammaglobulinemia.²⁴ A C-terminal His tagged construct of full length human BTK was expressed in baculovirus and purified using a nickel column. Purified protein was reduced in sample buffer (50 mM DTT, Pierce Rockford, IL, 90°C for 5 min) and alkylated (0.2 M iodoacetamide, Sigma, St Louis, MO) at room temperature for 20 min. The sample was separated on a 4–20% SDS-PAGE gel (Invitrogen, Carlsbad, CA). After fixation the gel was stained overnight in Coomassie Brilliant Blue and destained in 50% methanol. The gel band at 75kDa was excised and washed in 50 mM Ammonium bicarbonate in 50:50 acetonitrile: water. Gel pieces were dehydrated with acetonitrile and digested with trypsin (Promega, Madison, WI), in ammonium bicarbonate pH 8, 0.2 mg overnight at 37°C. Peptides were extracted from the gel in 50µL of 50:50 v/v acetonitrile: 1% formic acid (Sigma, St. Louis, MO) for 30 min followed by 50 µL of pure acetonitrile. Extractions were pooled and evaporated to near dryness and were reconstituted in 0.1% Formic acid. Samples were injected via an auto-sampler onto a 75 mM x 100 mm column (BEH, 1.7 mM, Waters Corp, Milford, MA) at a flow rate of 1 mL /min using a NanoAcquity UPLC (Waters Corp, Milford, MA). A gradient from 98% Solvent A (water + 0.1% Formic acid) to 80% Solvent B (acetonitrile + 0.08% Formic acid) was applied over 40 min. Samples were analyzed on-line via nanospray ionization into a hybrid LTQ-Orbitrap mass spectrometer (Thermo, San Jose, CA). Data was collected in data dependent mode

with the parent ion being analyzed in the FTMS at 60,000 resolution and the top 8 most abundant ions being selected for fragmentation and analysis in the LTQ.

Results

Mass Measurement Errors

Preview's measurements for the Plasma data set reveal randomly distributed rather than systematic errors, and recalibration makes only a small improvement, with the median precursor error improving from 508 to 370 ppm and the median fragment error from 0.076 to 0.072 Da. The Aurum data set shows systematic errors, but poor recalibration due to widely varying error characteristics over different MALDI plates. As shown in the Supplemental, the median precursor error actually grows worse with recalibration, but the larger precursor errors are improved. Preview's recalibration performs very well on Jurkat and, especially, BTK. On the Jurkat data set, recalibration improves the median precursor error from 7.2 ppm to 1.8 ppm and the median fragment error from 0.089 to 0.069 Da. On the BTK data set, recalibration improves the median precursor error from 15.4 ppm in BTK.dta to 2.3 ppm in BTK.recal.dta (Preview's output file) and the median fragment error from 0.285 to 0.104 Da. It is possible to obtain further improvement by submitting BTK.recal.dta as an input file to Preview and obtaining a new output file BTK.recal.recal.dta, with fourth-degree polynomial recalibration and median precursor error of 1.5 ppm and median fragment error of 0.068 Da. See Figure 2. More accurate mass measurements allow for a tighter matching tolerance, which leads to improved identification performance for almost any search program,²⁵ with slightly more true positives and substantially fewer false positives at the same score threshold. We recommend a mass tolerance at least three times the median accuracy reported by Preview, which is computed from the most intense, and hence the most accurate, peaks. On the Jurkat data set, for example, the optimal precursor tolerance (optimal by number of identifications) is 20 ppm with the original m/z measurements, and Byonic makes 2265 (fully tryptic, unmodified) identifications at an empirically estimated²⁰ False Discovery Rate (FDR) of 1%. With recalibrated measurements, the optimal tolerance is 10 ppm, about 5 times Preview's reported median error of 1.8 ppm, and Byonic makes 2350 identifications at 1% FDR, an improvement of about 4%. The identification improvement from recalibration of the BTK data set is much larger, about 20% more identifications at 1% FDR for Byonic, and about 15% more for X!Tandem.

Nonspecific Digestion and Modifications

Figure 3 shows that Preview's statistics on the prevalence of nonspecific cleavages and various modifications are in rough agreement with Byonic and X!Tandem on the Aurum, Jurkat, and Plasma data sets. The percentages in Figure 3 are all rates of modification, that is, the number of peptides with the property divided by the number that could have the property, for example, the number of peptide identifications including oxidized methionine divided by the number of methionine-containing peptide identifications. Since the three search engines make different identifications, the denominators in these rates are generally all different, yet none of them are so small as to be statistically insignificant. For example, Preview, which makes only about half as many identifications as the other search engines, has a 71/225 for pyro-glu, 44/167 for C[+71], and 39/184 for W[+32] on Aurum. Preview reports both fractions and percentages so that users can judge significance.

In Figure 3, Preview gave the measurement most different from the mean 5 times, X!Tandem 15 times, and Byonic 4 times, so the variability between Byonic and X!Tandem is at least as great as between Preview and the two conventional search engines, even though Preview limited attention to the top 100 proteins. Preview was run with no information

about modifications, not even the fixed cysteine modification. Byonic and X!Tandem require the user to specify which fixed and variable modifications to enable. In order to keep the conventional searches to a manageable size and avoid performance degradation, we did not enable modifications that through knowledge of the sample preparation method we knew not to expect: ST[+28] (formylation from exposure to formic acid) in Aurum; ST[+28], W[+32] (double oxidation to formylkynurenin), E[+14] (methyl esterification from methanol wash) in Jurkat; and MW[+32], E[+14], and C[+71] (cysteine propionamide, a gel artifact) in Plasma. Preview correctly found zero or near-zero levels for these modifications.

Figure 3 also reveals a stereotypical pattern for oxidation: highly oxidized samples such as Aurum have M[+16] in abundance and moderate amounts of M[+32], W[+32], H[+16], W[+16], and (not shown) C[+48]. Less oxidized samples have only M[+16] in quantity. Over 100s of data sets, Preview has shown that W[+32] and W[+16] co-occur, with the former almost always more common than the latter. Figure 3 shows mainly *in vitro* artifacts, but Preview also finds a low level of PTMs in the three data sets, for example, it finds three dimethylated lysines, one methylated asparagine, and one phosphorylated serine in Plasma. The phosphorylation, found in C[+58]DSSPDS[+80]AEDVR from Alpha-2-HS-glycoprotein, is known.²⁶

Preview's wild-card search finds an apparent N → S mutation in the peptide INN[−27]SLSELR (in protein IPI00001534 in Aurum), and indeed GenBank but not IPI contains both N and S forms of this protein. The wild-card search finds the peptide Y[−17]NSDLVQK in BTK; this is probably a post-source neutral loss of ammonia. It finds GEESSE[+16]MEQISIIR in BTK; this is probably GEESSEM[+16]EQISIIR with a noise peak causing a preference for E[+16]. The wild-card search also finds inexplicable mass shifts, for example, several occurrences of M[−46] in Aurum. Inexplicable mass shifts could be correct but unknown modifications (for example, we first discovered M[+34] by wild-card search and later identified it as homocysteic acid²⁷), sums of two or more modifications or mutations, or simply wrong answers.

Bruton's tyrosine kinase

Starting from mzXML (converted from RAW by ReadW), we compared two model data analysis approaches: (1) Byonic using expert knowledge but not Preview, and (2) Byonic with Preview for m/z recalibration and modification selection. We repeated the experiment using X!Tandem, with (1X) denoting X!Tandem without Preview and (2X) X!Tandem with Preview. We ran X!Tandem with all modifications enabled for both the first pass and the refinement.

For (1) and (1X), we chose a semitryptic search with any number of missed cleavages, with the following anticipated modifications: fixed C[+57] (alkylation), along with variable M[+16] (oxidation), STY[+80] (phosphorylation), and N-terminal Q[−17], E[−18], and C[+57][−17] (pyro-glu transformations). This benchmark is similar to Genentech's in-house analysis, but substituting Byonic and X!Tandem for Mascot. Mascot has a limit of about 7 modification types per search, so it was less suitable for this study than Byonic and X!Tandem.

For (2), we chose all the modifications that Preview found in any quantity, which included the anticipated modifications, along with C[+71] (propionamide), MW[+32] (oxidation), NQ[+1] deamidation, K[+14] (methylation), K[+28] (dimethylation), protein N-terminal +42 (acetylation), X[+22] (sodiation on any residue), and ST[−18] (beta elimination). For (2X) we included only the anticipated modifications, protein N-terminal +42, and ST[−18], because we found that X!Tandem's false and partially false discovery rates increase with a large number of modification types. We used 30 ppm precursor tolerance and 0.8 Da

fragment tolerance for (1) and (1X), and 20 ppm and 0.6 Da for (2) and (2X). We used a target / decoy database containing 583 proteins, including BTK, a few other kinases, and common contaminant proteins such as porcine trypsin and human keratins. Searches (1), (1X), (2), and (2X) were all moderately wide, with the large number of enabled modifications offset by the stringent precursor tolerances and small protein database.

Figure 4 shows the number of valid (top protein) identifications as a function of FDR. FDR was estimated as (# Reverse identifications) / (# of Forward identifications), running cumulatively down a list of spectrum identifications ranked by Byonic score, so FDR is not monotonically increasing. Preview-assisted Byonic produced new biological knowledge, phosphorylation sites²⁸ in BTK not found in current resources such as Uniprot Knowledge Base (www.uniprot.org/uniprot/Q06187), with higher confidence than Byonic alone. All 16 phosphorylations shown in Figure 4 are supported by multiple spectra and most by multiple distinct peptides, and manual curation changed none of Byonic's site localizations. We considered one localization ambiguous: the site is either S323 (as in Uniprot) or T324. In accord with the Paris Guidelines,²⁹ spectra for phosphopeptide identifications are given in the Supplemental.

Discussion

We designed the computational experiment on BTK to be a realistic demonstration of the use of Preview. For both the expert search and the Preview-assisted search we analyzed the data with a single Byonic or X!Tandem run, and did not perform a trial-and-error hunt for the optimal parameter settings. We did, however, build in one bias: we chose a data set that we knew to have atypically large m/z miscalibration. About 70% of the extra BTK identifications shown in Figure 3 came from enabling extra modifications and about 30% from m/z recalibration. Recalibration of m/z measurements is almost always beneficial, but the benefit may be negligible as in Plasma and Aurum, small as in Jurkat, or large as in BTK.

The most important unanticipated modifications were ST[−18], N[+1], X[+22], and protein N-terminal acetylation, with the other unanticipated modifications contributing only a few extra identifications. This experiment shows that modification setting is nontrivial, even with Preview. The optimal setting of modification parameters depends upon the prevalence of the modification; the cost of enabling the modification in search time, specificity, and sensitivity; and the goal of the study. As a general guideline, it makes sense to enable the most frequent modifications, reported on Preview's Summary page (see the Supplemental), along with all the PTMs that could shed light on the biology. Search time is generally a lesser consideration, but it is worth noting that some modifications (for example, protein N-terminal acetylation) cost almost no extra time, whereas others can be quite expensive (for example, sodiation on any residue).

In reanalysis of data sets from a wide variety of instruments and laboratories, we often find that the original data analysis lost sensitivity due to failure to consider some sample preparation artifact, most commonly non-specific digestion, over-alkylation, oxidation, and carbamylation. In some cases, the loss of sensitivity can be catastrophic, for example 90% loss of sensitivity at the protein level due to non-specific digestion in a sample of the fungus *Fusarium graminearum*³⁰. By providing a complete statistical overview of the data set before setting parameters for lengthy database searches, Preview can help improve both sample preparation and data analysis, ensuring that proteomics experiments yield maximum biological information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

MB was supported in part by NIH grants GM085718 and GM094557 and YJK by an NSF Computing Innovations postdoctoral fellowship.

References

1. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Electrophoresis. 1999; 20:3551–3567. [PubMed: 10612281]
2. Eng J, McCormack AL, Yates JR. J Am Soc Mass Spectrom. 1994; 5:976–989.
3. Craig R, Beavis RC. Bioinformatics. 2004; 20:1466–1467. [PubMed: 14976030]
4. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R. J Mass Spectrom. 2005; 40:430–443. [PubMed: 15838939]
5. Creasy DM, Cottrell JS. Proteomics. 2004; 4:1534–1536. [PubMed: 15174123]
6. Chamrad DC, Korting G, Stuhler K, Meyer HE, Klose J, Bluggel M. Proteomics. 2004; 4:619–628. [PubMed: 14997485]
7. Creasy DM, Cottrell JS. Proteomics. 2002; 2:1426–1434. [PubMed: 12422359]
8. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA. Mol Cell Proteomics. 2007; 6:1638–1655. [PubMed: 17533153]
9. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. Anal Chem. 2005; 77:4626–4639. [PubMed: 16013882]
10. Hernandez P, Gras R, Frey J, Appel RD. Proteomics. 2003; 3:870–878. [PubMed: 12833510]
11. Tang WH, Halpern BR, Shilov IV, Seymour SL, Keating SP, Loboda A, Patel AA, Schaeffer DA, Nuwaysir LM. Anal Chem. 2005; 77:3931–3946. [PubMed: 15987094]
12. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Nat Biotechnol. 2005; 23:1562–1567. [PubMed: 16311586]
13. Na S, Paek E. J Proteome Res. 2009; 8:4418–4427. [PubMed: 19658439]
14. Savitski MM, Nielsen ML, Zubarev RA. Mol Cell Proteomics. 2006; 5:935–948. [PubMed: 16439352]
15. Falkner JA, Falkner JW, Yocum AK, Andrews PC. J Proteome Res. 2008; 7:4614–4622. [PubMed: 18800783]
16. Bandeira N, Tsur D, Frank A, Pevzner PA. Proc Natl Acad Sci U S A. 2007; 104:6140–6145. [PubMed: 17404225]
17. Falkner JA, Kachman M, Veine DM, Walker A, Strahler JR, Andrews PC. J Am Soc Mass Spectrom. 2007; 18:850–855. [PubMed: 17329120]
18. Bern M, Cai Y, Goldberg D. Anal Chem. 2007; 79:1393–1400. [PubMed: 17243770]
19. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. Anal Chem. 2003; 75:4646–4658. [PubMed: 14632076]
20. Elias JE, Gygi SP. Nat Methods. 2007; 4:207–214. [PubMed: 17327847]
21. Boja ES, Fales HM. Anal Chem. 2001; 73:3576–3582. [PubMed: 11510821]
22. Chalkley RJ, Baker PR, Medzihradszky KF, Lynn AJ, Burlingame AL. Mol Cell Proteomics. 2008; 7:2386–2398. [PubMed: 18653769]
23. Davis RE, Ngo VN, Lenz G, Tolar P, Young RM, Romesser PB, Kohlhammer H, Lamy L, Zhao H, Yang Y, Xu W, Shaffer AL, Wright G, Xiao W, Powell J, Jiang JK, Thomas CJ, Rosenwald A, Ott G, Muller-Hermelink HK, Gascoyne RD, Connors JM, Johnson NA, Rimsza LM, Campo E, Jaffe ES, Wilson WH, Delabie J, Smeland EB, Fisher RI, Braziel RM, Tubbs RR, Cook JR, Weisenburger DD, Chan WC, Pierce SK, Staudt LM. Nature. 2010; 463:88–92. [PubMed: 20054396]

24. Blaese, RMWJ. Patient and Family Handbook for Primary Immunodeficiency Diseases. Immune Deficiency Foundation; Towson, MD: 2007.
25. Clauser KR, Baker P, Burlingame AL. *Anal Chem.* 1999; 71:2871–2882. [PubMed: 10424174]
26. Zahedi RP, Lewandrowski U, Wiesner J, Wortelkamp S, Moebius J, Schutz C, Walter U, Gambaryan S, Sickmann A. *J Proteome Res.* 2008; 7:526–534. [PubMed: 18088087]
27. Bern M, Saladino J, Sharp JS. *Rapid Commun Mass Spectrom.* 24:768–772. [PubMed: 20169556]
28. Oppermann FS, Gnad F, Olsen JV, Hornberger R, Greff Z, Keri G, Mann M, Daub H. *Mol Cell Proteomics.* 2009; 8:1751–1764. [PubMed: 19369195]
29. Bradshaw RA, Burlingame AL, Carr S, Aebersold R. *Mol Cell Proteomics.* 2006; 5:787–788. [PubMed: 16670253]
30. Padliya ND, Garrett WM, Campbell KB, Tabb DL, Cooper B. *Proteomics.* 2007; 7:3932–3942. [PubMed: 17922518]

Figure 1. Flowchart of Preview

The software performs a single initial pass over the full protein database, and then performs all subsequent searches on representative proteins and likely peptides from the representative proteins. It searches for modifications by performing successively wider searches, with the most likely modifications checked first in order to inform later searches.

Figure 2. Recalibration of mass measurements

Preview recalibrates both precursor and fragment m/z measurements based on confident peptide identifications. On the BTK data set, which has Orbitrap MS and LTQ MS/MS, the median precursor error (difference between observed and theoretical m/z) improves from 15.4 ppm to 2.3 ppm and the median fragment error from 0.285 to 0.104 Da, after correction with the quadratic recalibration curves shown in the upper panels. On this data set, it is possible to obtain further improvement to 1.5 ppm and 0.068 Da respectively, by resubmitting Preview's output spectrum file as a new input, finding a new and larger set of confident identifications (including fragment peaks above 1000 Da), and obtaining the recalibration curves shown in the lower panels.

Figure 3. Nonspecific digestion and prevalent chemical modifications

Preview, X!Tandem, and Byonic largely agree on percentages of semitryptic peptides (N- and Cragged) and prevalent modifications in Aurum, Jurkat, and Plasma. In all cases the percentage is the number of peptides with the property divided by the number that could have that property, for example, the percentage of methionine-containing peptides with at least one oxidized methionine.

Figure 4. Application to Bruton's tyrosine kinase

The ROC curves show the sensitivity/specificity tradeoff for X!Tandem and Byonic searches, with and without Preview. The boxes show phosphorylation sites in BTK found at 0%, 1%, and 2% FDR for the two Byonic searches; sites shown in bold such as Y223, the tyrosine residue at position 223, are already in UniProt KB. Without Preview, the MS/MS data was searched without m/z recalibration and with 7 anticipated modifications. With Preview, the data was searched with recalibration and up to 18 modifications.

Table 1

Shotgun proteomics data sets used in the study

We validated Preview's statistics by comparison with Byonic and X!Tandem on the first three data sets. Then we studied BTK (Bruton's tyrosine kinase) using Preview to guide parameter selection for Byonic and X!Tandem.

Name	Sample Type	Cysteine	Fractionation and Ionization	MS Instrument	Number of MS/MS scans
Aurum	247 reference proteins	C[+57], C[+71]	Gel purification+ MALDI	ABI 4700 TOF-TOF	9987
Jurkat	Jurkat cell lysate	C[+0]	Gel + LC-ESI	Thermo LTQ Orbitrap	9640
Plasma	Blood plasma	C[+58]	LC-ESI	Thermo LTQ	3824
BTK	Purified protein	C[+57]	Gel + LC - ESI	Thermo LTQ Orbitrap	3287