Determining the Overall Merit of Protein Identification Data Sets: *rho*-Diagrams and *rho*-Scores

David Fenyo,[†] Brett S. Phinney,[‡] and Ronald C. Beavis*[,§]

Rockefeller University, New York, New York 10019, University of California Davis, Davis, California 95616, and University of British Columbia, BC, Canada V6T 1Z3

Received January 16, 2007

This paper described a simple heuristic method for determining the merit of a set of peptide sequence assignments made using tandem mass spectra. The method involved comparing a prediction based on the known stochastic behavior of a sequence assignment algorithm with the assignments generated from a particular data set. A particular formulation of this comparison was defined through the construction of a plot of the data, the *rho*-diagram, as well as a parameter derived from this plot, the *rho*-score. This plot and parameter were shown to be able to readily characterize the relative quality of a set of peptide sequence assignments and to allow the straightforward determination of probability threshold values for the interpretation of proteomics data. This plot is independent of the algorithm or scoring scheme used to estimate the statistical significance of a set of experimental results; rather, it can be used as an objective test of the correctness of those estimates. The *rho*-score can also be used as a parameter to evaluate the relative merit of protein identifications, such as those made across proteome species taxonomic categories.

Keywords: rho-score • rho-diagram • bioinformatics • protein identification • peptide identification

1. Introduction

The determination of the proteins present in an experimental sample using tandem mass spectrometry has become an important technology in protein biochemistry and proteomics.[1] The most commonly used method for identifying the proteins present in a sample involves the use of enzymatic proteolysis to generate a set of peptides from the constituent proteins. These peptides are then separated from one another using one or more chromatographic steps, and then, the separated peptides are introduced into a tandem mass spectrometer and ionized. The masses of any peptides produced by the ion source are then measured, and each population of peptides is fragmented by collision with gas molecules or through a chemical reaction, generating a set of fragment ions that are dependent on the peptide's amino acid sequence. This operation of measuring fragment ion spectra is usually repeated for parent ion in the chromatogram that matches certain requirements, such as exceeding a sure-defined intensity threshold. A set of fragment ion mass spectra is generated that ideally should contain sufficient information to determine which proteins were present in the initial mixture. Over the course of the last 10 years, mass spectrometrists and chromatographers have worked to steadily improve the reliability of this data-taking exercise, resulting in instruments with the capacity to take more numerous (and more accurate) fragment ion mass spectra from a given protein sample.

At the same time, as the instruments for acquiring the data necessary to identify proteins has improved, a significant amount of progress has been made on the algorithms and software implementation necessary to reduce the raw tandem mass spectra into a set of correlations between these spectra and the list of protein sequences known to be potentially expressed by the organisms relevant to a particular study. These algorithms were developed based on the initial ideas associated with the automated annotation of fragmentation mass spectra[2−4] and an understanding of the sequence dependence of fragment ions generated by peptides.[5−6] These algorithms and software implementations (frequently referred to as "search engines"[7−10]) have made it possible to perform these "protein identifications" on very large data sets.

Protein identifications of this type are based on the assignment of one or more appropriate peptides from a protein's sequence to individual tandem mass spectra from the experimental data set. However, some fraction of the spectrum-to-peptide sequence assignments made by any algorithm will be caused by chance rather representing true sequence assignments. While these false-positive assignments can be easily dismissed by expert manual inspection in small sets of protein identifications, in very large sets, this approach has become impractical.[11]

Research into the solution of this problem has led to two general approaches. In the first approach,[12] an attempt is made to model the observed distribution of the best spectrum-to-

* To whom correspondence should be addressed at 2222 Health Science Mall, Vancouver, BC, Canada V6T 1Z3.

[†] Rockefeller University.

[‡] University of California Davis.

[§] University of British Columbia.

10.1021/pr070025y CCC: $37.00

Published on Web 03/31/2007

peptide sequence assignments in a large data set and fit this data to two theoretical distributions: a true-positive distribution and a false-positive one. The second approach[13] is to evaluate each spectrum-to-sequence assignment in the context of all possible sequence assignments for that spectrum, and by modeling the distribution of all assignments, determine the likelihood that the best assignment is significantly different from all other assignments. Both approaches use these model distributions to assign a probability that any particular best assignment could have occurred by chance alone.

These predictions of the probability that a particular sequence assignment was a chance event have been used to specify probability thresholds for interpreting a large data set. For example, a typical threshold value would be an expectation value $e < 0.05$, implying that the lowest quality assignment allowed would only be expected to occur 0.05 times at random (or once in 20 repetitions of similar experiments). These threshold values are normally set rather arbitrarily prior to the data analysis, and while they may produce reasonable results in general, using a single threshold value for data generated from different samples and different instruments may result in including an unnecessary number of false-positive assignments (too low a threshold) or false negatives (too high a threshold.

This paper proposes a simple method to solve the practical problem of how to determine a reasonable probability threshold value for a large data set and to easily evaluate the confidence that a particular data set represents true assignments in a single, easily obtained heuristic diagram. This type of diagram (*rho*-diagram) and a related parameter derived from it (*rho*-score) are the result of a simplification of any proteomics sequence assignment data set into a normalized plot designed to test the null hypothesis that all of the assignments present in a given data set are purely by chance. Inspection of the diagram allows the unambiguous assignment of reasonable threshold values, as well as a means of judging the degree to which changing any analysis parameter has improved (or worsened) the overall set of sequence assignments.

2. Definition of the *rho*-Diagram and *rho*-Score

The goodness-of-fit of a tandem mass spectrum to a peptide sequence is normally estimated using an algorithm that generates some type of score (s). The value of s is a function of some combination of factors that measure to what extent the spectrum matches the spectral features that can be predicted for that sequence. The scoring algorithms are designed to maximize s when the best correlations between a spectrum and a peptide are obtained.

Throughout the discussion below, the spectrum-to-sequence matches derived from this scoring process will be classified into two sets: true- or false-positives. The matches discussed correspond to the best possible assignments that can be made, following the most stringent application of algorithms and software designed to remove as many questionable matches as possible. These matches represent the results to be presented to the nonspecialist consumer of the information, such as biological or medical researchers, who have the reasonable expectation that any list of identifications reported to them corresponds to a list of positive results, albeit with varying degrees of confidence. The results of the much more detailed considerations associated with developing that final list, in which many more categories of assignments must be considered, can be tested using the heuristic described below, to the extent that the choices made impact the final list of identifications.

For the purposes of this discussion, a sequence assigned to a spectrum that is slightly different from the true sequence, but which was assigned because the true sequence was not tested, will be considered to be a true-positive. An example of this situation would be the assignment for which a detailed examination of the results showed that a better match to the data could have been obtained by assuming that a single asparagine residue in that sequence had been deamidated by the sample preparation protocol. On the other extreme, a strong spectrum that was assigned to a genuinely incorrect sequence because the true sequence was not present in the proteome sequences tested will be considered a false-positive. Any sequence assignment made to a spectrum generated from a nonpeptide precursor, such as chemical noise, will also be covered by the term false-positive.

Given the properties of s, it is possible to model the distribution of scores that any such algorithm will generate if it is applied to the comparison of a tandem mass spectrum and a set of peptides that does not contain any sequences that are true-positive identifications for that spectrum. This distribution of scores can be said to be stochastic, in the sense that it does not represent genuine identifications; it is generated by coincidental partial matches that are the result of comparing a spectrum against a large number of peptides sequences. The portion of the stochastic distribution ($p(s)$) representing the highest scores is of most interest in practice, as it is these high-scoring stochastic matches that are the most likely to be confused with true-positive. The functional form of $p(s)$ depends on the scoring algorithm, for example, forming a Poisson distribution[9] or a standard Gumbel distribution.[13] In most common cases, the region of the distribution representing the highest scoring stochastic matches can be approximated by an exponential distribution

$$p(s) \propto \exp(-\beta s) \tag{1}$$

If a particular data set contains many tandem mass spectra and these spectra have all been scored in such a way that they all represent stochastic matches (e.g., scoring them against a reversed protein sequence database[14]), then for a given score, it is straightforward to predict the number of spectra that would be expected to have a particular score, which is the definition of the expectation value, $e(s)$:

$$e(s) \propto p(s) \tag{2}$$

One way to test this relationship is to take a set of scored spectra and create a vector $\mathbf{E}$ composed of the number of spectra in intervals, such that E_x is defined as the number of spectra that have been assigned an expectation value between $\exp(-x)$ and $\exp(-(x + C))$. For simplicity, $C = 1$ will be used for this discussion (i.e., integer intervals) so that E_i ($i = 0, -1, -2, ,,,$) will be defined as the number of spectra with expectation values between $\exp(i)$ and $\exp(i - 1)$. Given a set of N spectra, each of which has been assigned a peptides sequence and an expectation value that the assignment is stochastic, if all of those assignments are truly stochastic, the values E_i can calculated from the definition of expectation values

$$E_i = \int_{\exp(i-1)}^{\exp(i)} N de = N[\exp(i) - \exp(i-1)] \tag{3}$$

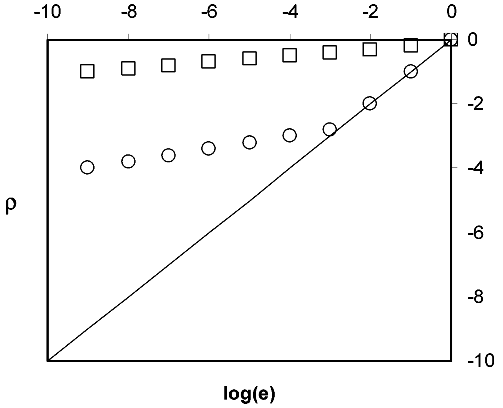

Figure 1. A representation of ρ as a function of the predicted log(e) distribution. The diagonal line represents purely stochastic matches, the squares (□) represent a hypothetical identification set that contains only true sequence identifications, and the circles (○) represent an intermediate hypothetical case containing both stochastic matches and true identifications.

A consequence of this definition of **E** is that the ratio of any two elements of this vector should have the following property:

$$\frac{E_{i-j}}{E_i} = \frac{N\exp(i-j)[1-\exp(-1)]}{N\exp(i)[1-\exp(-1)]} = \exp(-j), \quad \text{or}$$

$$\log\left(\frac{E_{i-j}}{E_i}\right) = -j \tag{4}$$

Equation 4 represents a prediction of the purely stochastic behavior of **E**, which can be very easily calculated for any set of experimental spectrum-to-sequence assignments. It should be noted that eq 4 is independent of eqs 1 and 2: it only requires that there is some method of estimating the expectation that a particular result is a stochastic sequence assignment. The estimator used to calculate the expectation value does not affect eq 4. Rather, the equation can be used to test whether a particular estimator is correctly assigning expectation values. Equation 4 is also independent of the scoring system used, the length of the peptides, the m/z value of the spectra, or any other experimental parameter. Any observed deviations from eq 4 that can be correlated with any such parameter indicate that the expectation value estimator is not adequately accounting for that parameter, making it a useful test for exploring the accuracy of any statistical formulation of the peptide-to-spectrum assignment problem.

A practical test of the null hypothesis presented by this eq 4 is to define a value $\rho(i)$, such that

$$\rho(i) = \log\left(\frac{E_i}{E_0}\right) \tag{5}$$

The utility of this value can be illustrated by constructing a plot of the measured $\rho(i)$ values for a particular set of spectrum-to-peptide identifications, as a function of the logarithm of the expectation value predicted for stochastic matches to the data. For the purposes of this discussion, a plot of this form will be referred to as a "*rho*-diagram".

Figure 1 illustrates several possible cases for a *rho*-diagram. If a set of putative peptide identifications is composed of purely stochastic matches, then the condition $\rho(\log(e(s)) = \log(e(s))$ will be met (the diagonal line). Alternatively, if a set of identifications represents purely true-positive identifications (open squares), then the intervals with low-expectation values will be populated with more matches than would be expected for the stochastic case, resulting in correspondingly higher ρ values. In the intermediate case, where the set contains both false-positive and true-positive identifications (open circles), the intervals with relatively high-expectation values are dominated by stochastic matches, but at some value of log(e(s)) true-positives begin to dominate, producing a deviation toward higher values for ρ. Simple visual inspection of this *rho*-diagram of any proteomics data set can therefore be used to rapidly evaluate the quality of the data set, as compared to purely stochastic behavior.

For quality control purposes, it is often useful to produce a single numerical value that can be used to classify the quality of a data set. One such value that can be derived from Figure 1 will be referred to as the "*rho*-score", which can be expressed by

$$\begin{aligned} r &= 100 \times (1 - R/D) \\ \text{where:} \quad & \text{if } r < 0, \quad \rho\text{-score} = 0 \\ & \text{if } 0 \le r \le 100, \quad \rho\text{-score} = r \\ & \text{if } r > 100, \quad \rho\text{-score} = 100 \end{aligned} \tag{6}$$

and where R is the integral of experimental data in the *rho*-diagram and D is the integral of the diagonal prediction, over the same range of log(e(s) values. Given this definition, the *rho*-score will range from 0 (indistinguishable from purely stochastic results) to 100 (putatively all true-positives). It should be noted that since both axes of this diagram are negative, these integrals represent the area between the curve and the negative x-axis, which is above the curve rather than below it.

3. Experimental Section

All protein identifications were performed using the open source search engine, X! Tandem (version 2007.01.01).[10] The *rho*-diagrams were generated by the Global Proteome Machine (GPM)[15] component PERL script "*plist.pl*" (version 2007.01.06). All of the software, including source code, was made available.[16] This version of X! Tandem generates the values for the vector **E**, in a range of log-intervals from $i = 0$ to -19, and stores them in its XML output files, in the "Performance parameters" group, as a note element with the attribute name "quality scores". The *rho*-values and diagrams reported in this paper were calculated by checking the E_i values, starting at $i = 0$ and stopping at the first $E_i < 5$.

It should be noted that the displays generated by the Global Proteome Machine interfaces use base-10 logarithms to display *rho*-diagrams, while all of the *rho*-diagrams shown in this work use natural logarithms. The functional form and conclusions formed from these diagrams was independent of the logarithm base used.

Protein identification analysis was performed using either an unmodified Panasonic Toughbook CF-W4 laptop computer (Pentium M, 1.2 GHz processor, 512 MB RAM) or an unmodified Sony Vaio PCV-W30 desktop computer (Pentium 4, 2.0 GHz processor, 512 MB RAM), using the standalone version of the GPM, gpm-xe.[16] All eukaryote proteome sequences were obtained from ENSEMBL.[17] All prokaryote proteome sequences were obtained from the National Center for Bioinformatics.[18]

The tandem mass spectrum data sets used for this analysis were obtained from data repositories, rather than generated especially for this study. Data sets from Proteome Commons,[19] the Open Proteomics Database,[20] and the PeptideAtlas Repository[21] were used, without any additional processing. Reference to the specific accession numbers and file names for specific data sets were made in the Results and Discussion section, below. Please refer to the original data annotations at the relevant repository to obtain information not provided in this manuscript regarding the details of how each data set was generated. These repositories also include the peptide-to-spectrum assignment information and protein identification information generated by the groups that made the original data depositions. Individual data sets were selected from these repositories based on their suitability for demonstrating features associated with *rho*-diagrams. The behavior of each of the selected data sets was not unusual: the features demonstrated below were common to all of the data sets available at these repositories tested.

These data sets fall into two categories: (1) data generated from artificial mixtures in which the sample composition are is well-known; and (2) data generated from real experiments, where the biological origin of the sample is known but the detailed composition of the sample is not. The Proteome Commons Aurum data set belongs to the first category: it was designed for use in testing protein identification algorithms. It consists of high-quality mass spectra obtained from a mixture of known proteins, with each of the recombinant proteins in the mixture having been purified individually and then mixed together in known concentrations. The data sets from OPD and PeptideAtlas belong to the second category: the data was obtained to solve real problems, and the types of experimental artifacts often found in real data are represented.

4. Results and Discussion

4.1. Testing the Accuracy of the Expectation Value Predictions. The derivation and justification for the *rho*-diagram given above assumes that the value of $e(s)$ for any spectrum-to-sequence match can be estimated with some degree of accuracy and that the distribution of $e(s)$ values for a data set will follow the behavior predicted by the appropriate stochastic model function. The search engine used in this case, X! Tandem, uses fit of the experimentally derived identifications to eq 1 as its estimator of stochastic behavior. While it is possible to construct theoretical data sets that would not conform to this model, the practical utility of *rho*-diagrams depends on the behavior of authentic data sets, generated by tandem mass spectrometers in the course of performing meaningful experiments. Therefore, all of the data sets used to test the assumptions made about the properties of *rho*-diagrams were obtained from publicly accessible proteomics experimental data repositories, rather than simulations or data specifically generated to demonstrate a particular point.

The validity of the X! Tandem's implementation of eq 1 for experimental data sets was demonstrated in Figure 2. This rho-diagram plots the results of performing protein identification analyses using protein sequences that do not contain any peptides that can be true spectrum-to-sequence assignments. These exclusively "false" protein sequences were generated by reversing the amino acid sequence of the proteins in the human proteome.[14] Therefore, the scores generated by this comparison were the result of purely false-positive stochastic assignments. The results of three independent data sets were plotted on the same diagram. Inspection of the diagram shows that, to a good degree of accuracy, the null hypothesis was correct. The plotted points clustered around the diagonal line, as would be expected given the prediction made in eq 4. The measured least-squares fit slope was 1.03, also in agreement with the predicted value of 1.00. On this basis, it was reasonable to conclude that eq 1 was an adequate expectation value estimator for these data sets. Many data sets and conditions for the generation of purely false-positive assignments have been tested in the course of this study (data not shown), and the plot in Figure 1 was representative of the results of these tests.

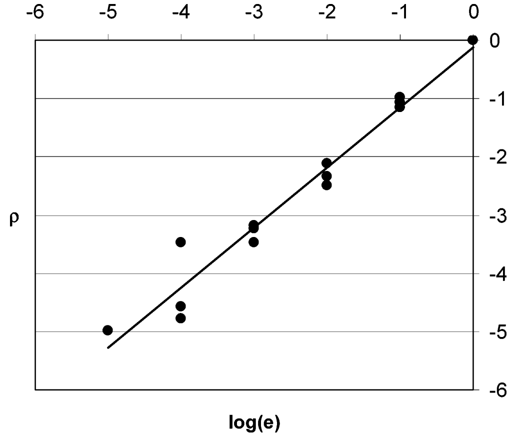

Figure 2. The *rho*-diagram generated from the analysis of three sets of experimental tandem mass spectra against reversed human protein sequences. The data sets used were obtained from the PeptideAtlas repository: PAe000032: *17.mzXML*; PAe000112: *AdducinIISCX1.mzXML;* and PAe000002: *raftapr_1.dta*. The solid line represents a least-squares fit to the data, resulting in $\rho = 1.03 \log(e) - 0.12$, $R^2 = 0.96$.

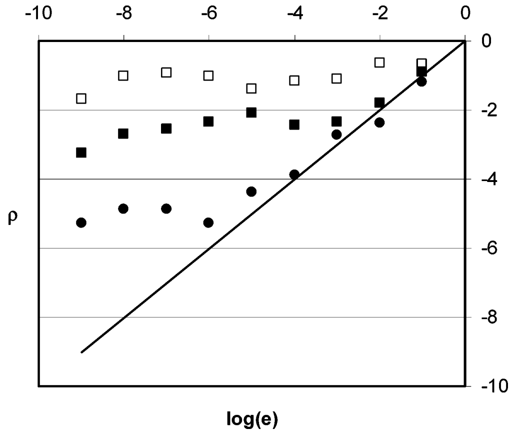

Figure 3. The *rho*-diagram generated from the analysis of two sets of experimental tandem mass spectra against human protein sequences. The data sets used were obtained from the PeptideAtlas repository: 1 (●, *rho*-score = 24) *PAe000032: 17.mzXML*; 2 (■, *rho*-score = 64) PAe000112: *AdducinIISCX1.mzXML;* and 3 Proteome Commons: (□, *rho*-score = 85) Aurum: *T10467.mgf*.

4.2. Interpretation of *rho*-Diagram Features. Figure 3 illustrated the *rho*-diagrams for three data sets, matched against the appropriate human protein sequences. Data set 1 (closed circles) was characteristic of the *rho*-diagram generated from a spectrum collection of rather low quality, containing few interpretable spectra. The points on the plot began to deviate significantly from the diagonal beginning at $\log(e) \approx -6$ (or $e(s) \approx 0.0025$). The interpretation of this behavior was understood by thinking of the assigned spectra in each interval as being a mixture of true-positive (t_i) and false-positive assignments (f_i). From the definition of **E**:

$$E_i = t_i + f_i \tag{7}$$

Therefore, for Data set 1, in the region where the points were on the diagonal, $f_i \gg t_i$ and the behavior of r was dominated by the behavior of f_i, which was governed by the stochastic distribution. Deviation from the diagonal begins when $t_i \approx f_i$. In the region of the curve where $t_i \gg f_i$, the behavior of the plot became independent of the stochastic distribution; instead the plot followed the distribution of true-positive results. It should be remembered that any useful scoring algorithm has been specifically designed to generate high true-positive scores so it should generate a distribution significantly skewed toward high scores.

This interpretation allows one to directly imply that for results in Data set 1 with $\log(e(s)) > -5$, a majority of the peptide sequence assignments were produced by stochastic matches, and therefore the results are unreliable. Results with $\log(e(s)) < -7$ were nonstochastic and represent true-positive assignments. Results in the range $\log(e(s)) > -7$ but < -5 contain a mixture of a similar number of true and false-positive assignments. It should be emphasized that this interpretation does not mean that there are no true-positive assignments among the results with $\log(e(s)) > -5$. Rather, it means that it would be necessary to apply additional constraints during the analysis process to distinguish the relatively rare true-positives from the predominantly false-positive background.

Data set 2 (Figure 3, closed squares) represented a more common case, and it was typical of most good quality data sets examined for this study. The transition between stochastic and true-positive behavior was sharper than for Data set 1, occurring at $\log(e(s)) \approx -3$ (or $e(s) \approx 0.05$). Results with $\log(e(s)) < -3$ would be dominated by true-positive peptide sequence assignments, while those with $\log(e(s)) > -3$ would be dominated by false-positive ones.

Data set 3 (Figure 3, open squares) represented the best case, and it was typical of the highest quality data sets examined. In this case, there were no data points that could be convincingly fit to the diagonal line. Conservatively, results with $\log(e(s)) < -1$ (or $e(s) < 0.4$) were interpreted as being mainly composed of true-positives: for this data set, there was no convincing evidence that any significant number of false-positive results were found.

The behavior of the plots displayed in Figure 3 made their interpretation in terms of a threshold (e_{thresh}) quite simple. In normal laboratory usage, this threshold represented an expectation value such that for all $e(s) < e_{\text{thresh}}$, a majority of the peptide assignment were true-positives. In terms of the *rho*-diagram, this condition was met when the experimental points began to deviate systematically from the diagonal line. By projecting that point of deviation up on to the x-axis, the $\log(e(s))$ value corresponding to the change from majority false-positive to majority true-positive was estimated. Application of this idea to Figure 3 yielded Data set 1, $\log(e_{\text{thresh}}) = -6$; Data set 2, $\log(e_{thresh}) = -3$; and Data set 3, $\log(e_{\text{thresh}}) = -1$.

Consideration of all three data sets in Figure 3 together demonstrated the potential practical advantages of using a *rho*-diagram to determine the probability threshold for the interpretation of proteomics data. Simply selecting a single value for this threshold might be adequate for one of the sets, but it would be completely inadequate for the other two, either including too many false-positives or excluding too many true-positives. Determining the threshold value by directly testing the stochastic null hypothesis, however, produces a very reasonable heuristic criterion for establishing its value as well as a simple test that allows for the rejection of any data set that does not contain a significant number of interpretable sequence assignments.

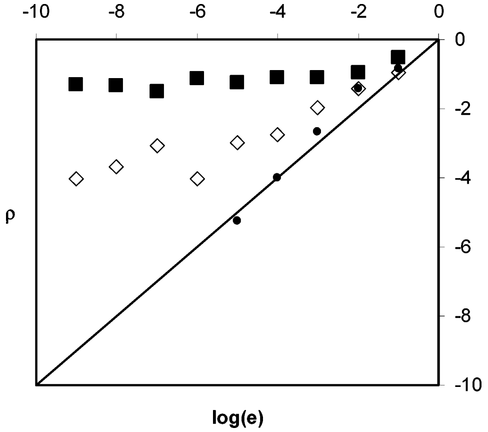

Figure 4. Demonstration of the use of the *rho*-diagram to evaluate the effects of changing the parameters for a search engine, using the PeptideAtlas experimental tandem mass spectra data set PAe000166: *lb031104_04.mzXML*, searched against the *S. cerevisiae* proteome. The relevant parameter sets were as follows: 1 (■, *rho*-score = 84) parent ion mass tolerance +4/−0 Da, protein cleavage reagent: trypsin; 2 (◇, *rho*-score = 44) parent ion tolerance −0/−4 Da, protein cleavage reagent: trypsin; and 3 (●, *rho*-score = 0) parent ion mass tolerance +4/−0.5 Da, protein cleavage reagent: V8 protease.

An important special case occurred when a particular result set generated a *rho*-diagram in which there was no significant deviation from the diagonal (e.g., all three data sets in Figure 2). In this case, all of the intervals in **E** contained a majority of false-positive results. The concept of a threshold value was inappropriate for analyzing this type of results set, and the results must be either be interpreted manually or rejected as unreliable and remeasured.

4.3. Using *rho*-Diagrams To Optimize Search Parameters. Any practical search engine uses a number of parameters to customize the sequence assignment process to the characteristics of the chemical processing and instrumentation used to generate and measure the peptide tandem mass spectra necessary for the sequence assignment process. Determining the appropriate values for these parameters has been a matter of intuition and experience on the part of the analyst.

Figure 4 illustrates the use of a *rho*-diagram to evaluate the results of changing these search parameters. The closed rectangles plot the results of setting the parent ion tolerance in the range $4 > \Delta m > 0$ Da ($\Delta m = m - M$, where m was the measured parent ion mass, and M was the mass calculated from the peptide sequence). The open diamonds plot the results of searching the same data set, with $0 > \Delta m > -4$ Da. Clearly, the plots were quite different: the first parameter setting produces a plot that showed no evidence of significant false-positives in any interval, while the second setting produces a plot that only slowly separates itself from diagonal. This behavior indicated that the second setting produced far fewer true-positives than the first. This effect was predictable: the low resolution of the quadrupole mass spectrometer used to make the measurements has the practical effect of interpreting the contribution of ^{13}C isotopes as an effective mass shift to higher measured mass. The additional effects of space-charging in the ion trap also tended to shift the measurement toward higher mass. Therefore, most true parent ion mass measurements should be higher than the calculated mass. A small number of true measurements had $\Delta m < 0$; however, their unambiguous interpretation would clearly require more effort.

The third plot in Figure 4 (closed circles) was generated by using a parent ion tolerance that should be good ($4.0 > \Delta m > -0.5$ Da); however, the protein cleavage reagent parameter was changed from cleavage with trypsin (which cleaves at any K−X or R−X bond, except for X═P) to cleavage with V8 protease (which cleaves at any E−X or D−X bond). The parameter setting produced a diagonal plot (least-squares fit linear slope = 1.05) indicating that the data contained no evidence of true-positive results. The cleavage reagent used in the actual experiment was trypsin, so the use of any cleavage parameter mutually exclusive to trypsin's sequence specificity would be expected to produce only stochastic results.

4.4. Evaluation of Results Using the *rho*-Score. The discussion of Figures 2−4 has stressed the functional form of the plotted data, and utility of these diagrams as an unbiased diagnostic tool for understanding any large set of proteomics results. Many common circumstances exist in the practice of laboratory proteomics where examining the details of a *rho*-diagram would not be necessary. Instead, comparing results based on the simpler *rho*-score (eq 6) may be sufficient. This score was designed so that its interpretation would be simple: the higher the *rho*-score, the greater the proportion of true-positives in any result set.

When this approach was used to analyzing the data in Figure 3, it was simple to order the true-positive content of the results: Data set 3 (*rho*-score = 85) > Data set 2 (*rho*-score = 64) > Data set 1 (*rho*-score = 24). From the view point of the scientists engaged in performing this type of analysis, characterizing these data sets by a single, normalized numerical value considerably simplifies the task of expressing the relative quality of data sets. From the view point of the consumer of proteomics information (e.g., a biologist attempting to evaluate proteomics data reported in the literature), the existence of a normalized figure-of-merit of this type may simplify the task of drawing conclusions.

The results in Figure 4 may also be ordered in terms of their true-positive content: analysis no. 1 (*rho*-score = 84) > analysis no. 2 (*rho*-score = 44) > analysis no. 3 (*rho*-score = 0). While it may have been possible to determine this ordering in other ways, for example, by performing a detailed examination of the results, the use of a single score with a straightforward statistically valid implication greatly reduces the expertise necessary to make such a judgment. Proteomics data has often been analyzed to maximize the number of identifications found, with no method available to determine how much of any increase in this number of identifications was caused by the increasing proportion of false-positive results. The use of *rho*-scores as an additional criterion may provide insight into this latter mechanism, which can confound the application of any optimization process to real experimental data.

4.5. Use of the *rho*-Score for the Comparison of Cross-Species Proteomics Results. One of the practical problems associated with the use of tandem mass spectra and peptide identification search engines that compare the spectra to the sequences from a known proteome has been the lack of known proteome sequence sets for some commonly used experimental model organisms. While this situation has been resolved for many organisms because of the increasing number of fully sequenced genomes, it still remains a challenge. For example, if an investigator has thoroughly characterized a Syrian hamster (*Mesocricetus auratus*) model for a particular form of cancer, to what extent can the well-known proteome sequence of the house mouse (*Mus musculus*) be used to interpret a proteomics data set derived from that hamster model? Would the sequence of the Norway rat (*Rattus norvegicus*) be a better choice?

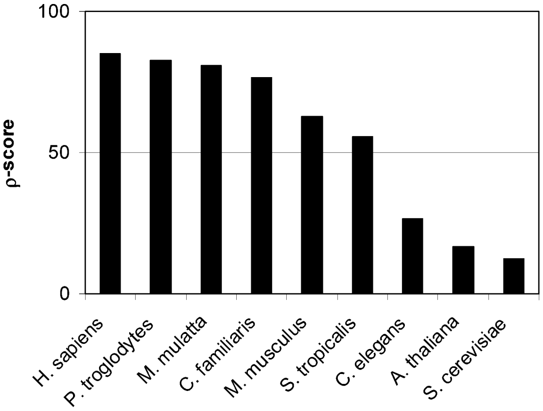

Figure 5. The *rho*-scores generated from the analysis of the complete Proteome Commons Aurum spectrum data set of experimental tandem mass spectra against different eukaryote species proteome sequences. The source species of the sample was *H. sapiens*.

Given a set of experimental spectra and a list of sequences for each species, the application of *rho*-scores as a figure-of-merit can provide at least some guidance to answer this type of question. Figure 5 displayed the comparison of *rho*-scores generated by searching a large set of high-quality tandem mass spectra against nine different eukaryote species' proteomes. The experimental data was generated from a set of human-sequence proteins. The trend shows that species closely related to humans, for example, *Pan troglodytes* (chimpanzee) and *Macaca mulatta* (rhesus macaque), produce *rho*-scores closest to the human value. These close relatives are followed (in order of decreasing true-positive content) by the dog, mouse, frog, nematode, mouse ear cress, and brewer's yeast proteomes. Given this result, using the chimpanzee or rhesus macaque proteome to analyze human-sourced proteins should give reasonably good results. Using the brewer's yeast proteome would be a bad choice.

The temptation to use this type of correlation to draw taxonomic conclusions about eukaryote species should be treated with some caution. An experimental data set generated from niche-specific proteins may produce a very different result from one generated from proteins that have highly conserved sequences. Differences in the quality of genome sequences may also influence the species-to-species trends for particular data sets. A species with a low coverage genome may be the missing the necessary coding regions for the translation of a particular protein because genome sequence is incomplete, rather than the absence (or significant evolution) of that protein sequence.

It may also be tempting to equate the *rho*-score value with other commonly used characterizations of this type of data set, for example, assuming that a *rho*-score of 25 implies that the results were 25% correct. While there will be correlations between different figures-of-merit for a data set, the *rho*-score should only be considered as a relative figure-of-merit to compare differences between sets of results and not as a fractional measure of "correctness" of any particular result.

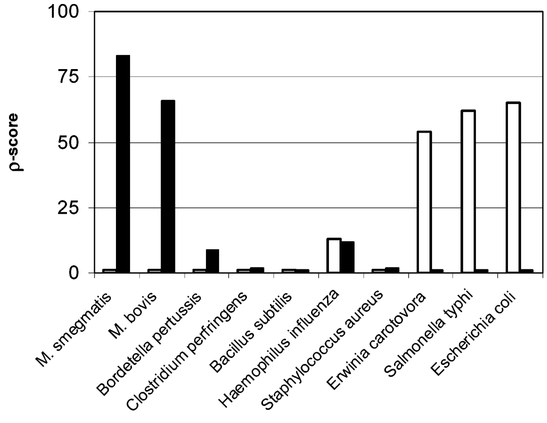

Figure 6. The *rho*-scores generated from the analysis of two sets of experimental tandem mass spectra from the Open Proteome Database, OPD00001_ECOLI (*E. coli* sample, open bars) and OPD00009_MYCSM (*M. smegmatis* sample, solid bars) analyzed using different prokaryote species proteome sequences.

A similar examination of prokaryote species was illustrated in Figure 6. Two different sets of spectra, one derived from *Escherichia coli* proteins and the other from *Mycobacterium smegmatis* proteins, were scored against 10 species' proteome sequences. Again the trends in the *rho*-score mirror the expected taxonomic relationships, although in this case the signals for unrelated species were very low. From this study and others performed on other data sets (not shown), only species that were closely related to the species that was the source of the proteins produced significant *rho*-scores. Exceptions to this generalization will certainly occur, with similar caveats to those made above for eukaryotes. The smaller number of proteins in a prokaryote's proteome and the capacity of prokaryotes to evolve more rapidly than multicellular eukaryotes may also explain the relatively sharp distinctions between species revealed by this plot.

Figure 6 also illustrates the capability of this simple analysis to signal small weaknesses in an expectation value estimation algorithm. The *Haemophilus influenza* proteome produced a *rho*-score ~10 for both sets of data, while most of the other unrelated proteomes produced a *rho*-score ~0. While a *rho*-score of 10 was clearly not as good a match as those obtained from truly related species, it was different from the other nonrelated species. This difference can be attributed to the way that the expectation value estimator used by X! Tandem works, rather than any taxonomic relationship between the species. X! Tandem tracks the distribution of scores for all possible peptide sequences from a target proteome that can be made to a particular spectrum and uses that distribution to fit eq 1. Because *H. influenza* has the smallest known bacterial proteome (one-third the size of *E. coli*'s), X! Tandem's expectation value calculation would be expected to be somewhat less accurate for that proteome, resulting in the observed small relative increase in the *rho*-score. To make the best use of X! Tandem for a proteome as small as *H. influenza*'s, it would be advisable to add sufficient decoy sequences into the protein list so that the total number of proteins searched would be the equivalent to that of a more typical bacterial proteome (3000–4000 proteins).

Conclusions

A novel plot, the *rho*-diagram, for testing proteomics results sets has been demonstrated to have some utility as a heuristic method of evaluating the overall quality of the results. The diagram was derived from the null hypothesis that all of the spectrum-to-sequence assignments generated from a set of experimental tandem mass spectra were false-positives caused by stochastic matches between the spectra and candidate peptide sequences. The validity of the modeling of distributions describing these stochastic matches was verified, and the general characteristics of this diagram were illustrated by examples drawn from public data repositories. A score derived from this plot, the *rho*-score, was shown to reflect the extent to which a proteomics result set was composed of true-positive sequence assignments. This score was shown to have potential use for optimizing the parameters used to carry out spectrum-to-sequence assignments. This score may also be useful in evaluating the merit of analyzing experimental proteomics data with the proteome sequences of taxonomically related species that have fully sequenced genomes.

Acknowledgment. The authors would like to thank Juergen Kast for his helpful comments on the manuscript. We would also like to thank Peter Lobel and John Wilkins for their efforts in testing the software implementation of the ideas in this paper and their comments on its usefulness. This work was supported, in part, by funding provided by the Biological Research Centre at the University of British Columbia, Beavis Informatics, Ltd., and the U.S. National Cancer Institute's Clinical Proteomics Technology for Cancer program, Grant Number 1 U24 CA126477-01.

References

(1) Aebersold, R.; Goodlett, D. R. Mass spectrometry in proteomics. *Chem. Rev.* **2001**, *101*, 269–295.

(2) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. Applications of artificial intelligence for chemical inference. I. The number of possible organic compounds. Acyclic structures containing C, H, O, and N. *J. Am. Chem. Soc.* **1969**, *91*, 2973–2976.

(3) Duffield, A. M.; Robertson, A. V.; Djerassi, C.; Buchanan, B. G.; Sutherland, G. L.; Feigenbaum, E. A.; Lederberg, J. Applications of artificial intelligence for chemical inference. II. Interpretation of low-resolution mass spectra of ketones. *J. Am. Chem. Soc.* **1969**, *91*, 2977–2981.

(4) Schroll, G.; Duffield, A. M.; Djerassi, C.; Buchanan, B. G.; Sutherland, G. L.; Feigenbaum, E. A.; Lederberg, J. Applications of artificial intelligence for chemical inference. III. Aliphatic ethers diagnosed by their low-resolution mass spectra and nuclear magnetic resonance data. *J. Am. Chem. Soc.* **1969**, *91*, 2977–2981.

(5) Roepstorff, P.; Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **1984**, *11*, 601.

(6) Biemann, K.; Martin, S. A. Mass spectrometric determination of the amino acid sequence of peptides and proteins. *Mass Spectrom. Rev.* **1987**, *6*, 1–76.

(7) Eng, J.; McCormack, A.; Yates, J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976.

(8) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–2567.

(9) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958–964.

(10) Craig, R.; Beavis, R. C. X! TANDEM: matching proteins with mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.

(11) Rappsilber, J.; Mann, M. What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **2002**, *27*, 74–78.

(12) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383 –5392.

(13) Fenyö, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75*, 768−774.

(14) Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2005**, *2*, 667−675.

(15) Craig, R.; Cortens, J. P.; Beavis, R. C. An open source system for analyzing, validating and storing protein identification data. *J. Proteome Res.* **2004**, *3*, 1234−1242.

(16) ftp://ftp.thegpm.org.

(17) ftp://ftp.ensembl.org/pub.

(18) ftp://ftp.ncbi.nih.gov/genomes/Bacteria.

(19) Falkner, J. A.; Falkner, J. W.; Andrews, P. C. ProteomeCommons.org JAF: reference information and tools for proteomics. *Bioinformatics* **2006**, *22*, 632−633.

(20) Prince, J. T.; Carlson, M. W.; Wang, R.; Lu, P.; Marcotte, E. M. The need for a public proteomics repository. *Nat. Biotechnol.* **2004**, *22*, 471−472.

(21) Desiere, F.; et al. Integration of peptide sequences obtained by high-throughput mass spectrometry with the human genome. *GenomeBiology* **2004**, *6*, R9.

PR070025Y