nature biotechnology

Intensity-based protein identification by machine learning from a library of tandem mass spectra

Joshua E Elias[1], Francis D Gibbons[2], Oliver D King[2], Frederick P Roth[2,4] & Steven P Gygi[1,3,4]

Tandem mass spectrometry (MS/MS) has emerged as a cornerstone of proteomics owing in part to robust spectral interpretation algorithms[1–6]. Widely used algorithms do not fully exploit the intensity patterns present in mass spectra. Here, we demonstrate that intensity pattern modeling improves peptide and protein identification from MS/MS spectra. We modeled fragment ion intensities using a machine-learning approach that estimates the likelihood of observed intensities given peptide and fragment attributes. From 1,000,000 spectra, we chose 27,000 with high-quality, nonredundant matches as training data. Using the same 27,000 spectra, intensity was similarly modeled with mismatched peptides. We used these two probabilistic models to compute the relative likelihood of an observed spectrum given that a candidate peptide is matched or mismatched. We used a 'decoy' proteome approach to estimate incorrect match frequency[7], and demonstrated that an intensity-based method reduces peptide identification error by 50–96% without any loss in sensitivity.

We constructed a training data set of high-confidence peptide-spectrum matches (PSMs) for modeling peptide fragmentation within ion-trap mass spectrometers (**Supplementary Figs. 1,2** online). Multiple protein sources were analyzed on electrospray ion-trap mass spectrometers to generate one million MS/MS spectra. Of these, roughly 140,000 PSMs were selected with high confidence because the top SEQUEST[8]-ranked peptide exceeded scores shown previously to yield an error rate of <1%[7]. We restricted our analysis to doubly charged peptides with two tryptic termini (fully tryptic) for this proof-of-concept experiment. This peptide class is estimated to represent more than 70% of correctly matched peptides[7]. Redundant PSMs were eliminated, reducing overrepresentation of abundant proteins. Analysis of redundant PSMs (**Supplementary Fig. 3** online) shows that fragment intensities are reproducible, a minimal requirement for our proposed method. We trained our algorithm on the remaining 27,266 spectra.

We implemented a probabilistic decision tree[9] to model the probability of observing a fragment ion intensity, conditioned on 63 peptide and fragment attributes such as those listed in **Table 1**, using the PSMs above as training data. Reduction in the Shannon entropy of intensity was used to select attributes and corresponding values for decision points[10,11]. Tree branches were terminated according to the Bayesian information criterion, previously used with decision trees to avoid over-fitting[10].The resulting match tree is graphically represented in **Figure 1**. A mismatch tree, trained on incorrectly matched PSMs, is shown in **Supplementary Figure 4** online.

The root node of each tree shows the attribute yielding the most information about fragment ion intensity within the entire training set, that is, the attribute providing the greatest expected reduction in entropy. For both match and mismatch trees (**Fig. 1** and **Supplementary Fig. 4** online) this is "POS > 0.16," indicating that ions derived from fragmentation within the 16% of a peptide's length closest to the N terminus nearly always yield low-intensity MS/MS peaks. Attributes and their abbreviations are described in **Table 1**. Subsequent internal nodes (ellipses) represent attribute and value combinations that best segregate the resulting subgroup of fragment ions. Terminal ('leaf') nodes are labeled with the number of training set fragments assigned to the node and reflect the intensity distribution of these fragments. Internal and leaf nodes are color coded according to the intensity range, or bin, that was most frequently observed (**Fig. 1**).

The decision tree automatically rediscovered qualitative phenomena known previously to experienced mass spectrometrists. As represented by the first two nodes (**Fig. 1**), fragmentation events near the N or C termini are most likely to be observed with low intensities, if at all. Also, as indicated by branches below the ION node, *y*-type ions tend to be more intense (primarily blue-violet) than *b*-type ions. Furthermore, it is known that proline strongly directs fragmentation of its N-terminal peptide bond whereas cleavage of its C-terminal bond is reduced[12–14]. The influence of this residue is evident throughout the match tree (RESN_1 = P, RESC_1 = P).

The tree also identified previously undescribed rules. For example, proline's inhibitory effect on fragmentation may extend to the second C-terminal peptide bond (RESC_2 = P). Furthermore, attributes describing residues N-terminal to fragmentation sites (GBN_1, HLXN_1, HYDN_1) are prevalent whereas similar attributes describing residues C-terminal to fragmentation sites are absent. This suggests that with the exception of proline, fragmentation may

[1]Department of Cell Biology, [2]Department of Biological Chemistry and Molecular Pharmacology, [3]Taplin Biological Mass Spectrometry Facility, 250 Longwood Avenue, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]These authors contributed equally. Correspondence should be addressed to F.P.R. (fritz_roth@hms.harvard.edu).

Published online 18 January 2004; doi:10.1038/nbt930

Table 1 Abbreviations and descriptions for attributes appearing in match and mismatch decision trees

Tree abbreviation	Attribute description	Tree
DISTC	Distance (number of residues) from fragmentation site to C terminus	Match, mismatch
DISTN	Distance (number of residues) from fragmentation site to N terminus	Match, mismatch
FRACF_HKR	Fraction of histidines, lysines, arginines in fragment ion	Match
GBC_X	Gas phase basicity of residue X positions C-terminal to fragmentation site[19]	Mismatch
GBF	Average gas phase basicity of fragment ion[19]	Mismatch
GBN_X	Gas phase basicity of residue X positions N-terminal to fragmentation site[19]	Match
HLXN_X	Helicity of residue X positions N-terminal to fragmentation site[20]	Match
HYDN_X	Hydrophobicity of residue X positions N-terminal to fragmentation site[20]	Match
ION	*b*- or *y*-type ion	Match
LENF	Length of fragment ion	Mismatch
LENP	Length of peptide	Match
M_Z	Fragment *m/z*	Match, mismatch
NTERM	Identity of residue at peptide's N terminus	Match
NUMP_HKR	Number of histidines, lysines, arginines in peptide	Match
PMASSD	Fragment mass – precursor mass	Match, mismatch
PMZ	Precursor ion *m/z*	Match
PMZD	Fragment *m/z* – precursor *m/z*	Match, mismatch
POS	Fractional location of fragmentation site along peptide	Match, mismatch
RESC_X	Identity of residue X positions C-terminal to fragmentation site	Match
RESN_X	Identity of residue X positions N-terminal to fragmentation site	Match

A complete attribute list is presented in **Supplementary Table 3** online.

be related more to properties of the N-terminal residue than to those of the neighboring C-terminal residue. The automated discovery of known rules that qualitatively predict fragment intensity lends credence to our approach and supports the validity of these and other novel rules discovered by the decision tree.

Rigorous filtering criteria gave us confidence that each training set spectrum was correctly matched to SEQUEST's top-ranked peptide. Lower-ranked peptides, though their predicted MS/MS spectra may resemble the observed spectrum, are therefore likely to be incorrect. We trained a mismatch probabilistic decision tree using the same training spectra matched with second-ranked peptides (**Supplementary Fig. 4** online). This mismatch tree serves as a control to the match tree, permitting comparison of the alternative hypotheses that a peptide is correctly or incorrectly matched. The second-ranked match was chosen over a randomly selected peptide so that alternative PSMs scoring similarly by existing methods will be discriminated from one another. Such discrimination represents a dilemma often faced in interpreting MS/MS spectra.

Taking the predicted fragment ions from a peptide sequence as input, one can obtain the likelihood of observing each measured intensity under both match and mismatch models. Although a probability score relying solely on the match model might discriminate between correct and incorrect PSMs, such a score—involving a product of fragment intensity likelihood terms, each less than unity—would unfairly penalize long peptides. We circumvent this problem with a log-odds ratio (*lod*) approach. The odds ratio for the i^{th} fragment refers to the ratio of likelihoods (p) of the observed fragment intensity under the alternative hypotheses of a correctly (*match*) or incorrectly (*mismatch*) matched candidate peptide (**Supplementary Fig. 5** online):

$$lod_i = \log_{10}\left(\frac{p(observed\ intensity \mid attributes,\ match)}{p(observed\ intensity \mid attributes,\ mismatch)}\right)$$

Summing lod_i scores over all (N_f) predicted fragments, we obtain the overall *LOD* score for a peptide:

$$LOD = \sum_{i=1}^{N_f} lod_i$$

Positive *LOD* scores suggest the peptide is more likely to be correctly than incorrectly matched.

Supplementary Figure 5 online shows how match and mismatch trees may be consulted to generate a lod_i score for a single predicted fragment ion. Performing this procedure on all predicted fragments from the peptide SALSGHLETLILGLLK (not used in training) yields a lod_i spectrum (**Fig. 2**). This spectrum reveals that although some fragments' intensities are more likely to have arisen from the mismatch probability distribution (*e.g.*, *y4*, *y5* and *y6*), observed intensities for most predicted ions have positive lod_i scores. The *LOD* score 2.22 suggests that this spectrum is 160 times more likely to be correctly matched than mismatched (that is, to have arisen from the match PSM model than the mismatch PSM model). In comparison, the second SEQUEST-ranked peptide for this spectrum, ISADFHVDLNHAAVR, received a *LOD* score of –3.05, indicating this spectrum is 1,000 times more likely to correspond to the mismatch PSM model. Although both peptides received relatively high *XCorr* (see **Table 2**) scores (2.7696 and 2.5580), the relative correlation difference between them, measured by the ΔCn (see **Table 2**) score (0.0764), is insufficient to permit confident selection of either peptide by most published criteria. This example suggests that the intensity-based *LOD* score provides additional useful information for correct PSM selection. Moreover, lod_i spectra such as those in **Figure 2** may be useful in directing attention to predicted fragment ions incongruent with the observed MS/MS spectrum. An example of *LOD* score performance for a single protein isolated by SDS-PAGE is presented in **Supplementary Table 1** online.

To rigorously investigate whether the *LOD* score distinguishes correct from incorrect PSMs on large-scale data, we used a decoy proteome strategy (Methods). **Figure 3a** illustrates the relationship between precision and sensitivity for the scores described in **Table 2**. An optimal score would achieve both 100% sensitivity and precision and would be represented as a point in the upper right corner of this graph. *LOD* outperforms the other three single-component scores (*XCorr*, ΔCn, ΔLOD, defined in **Table 2**) over the full range of precision and sensitivity. Furthermore, the two composite scores (*Disc*, *CompLOD*, defined in **Table 2**) consistently outperform their constituent single-component scores[15]. This analysis excluded spectra with matched peptides used in training. A repeated analysis including these spectra indicated greater performance for all scoring methods, but essentially the same relationships between scoring methods (data not shown).

LOD-based scores are complementary to the traditional scores *XCorr* and ΔCn. Combinations between these two scoring methods, such as *LOD* with *XCorr,* outperformed combinations within each method, such as *XCorr* with ΔCn or *LOD* with ΔLOD (data not shown). Similarly, *CompLOD,* which incorporates *LOD*-based

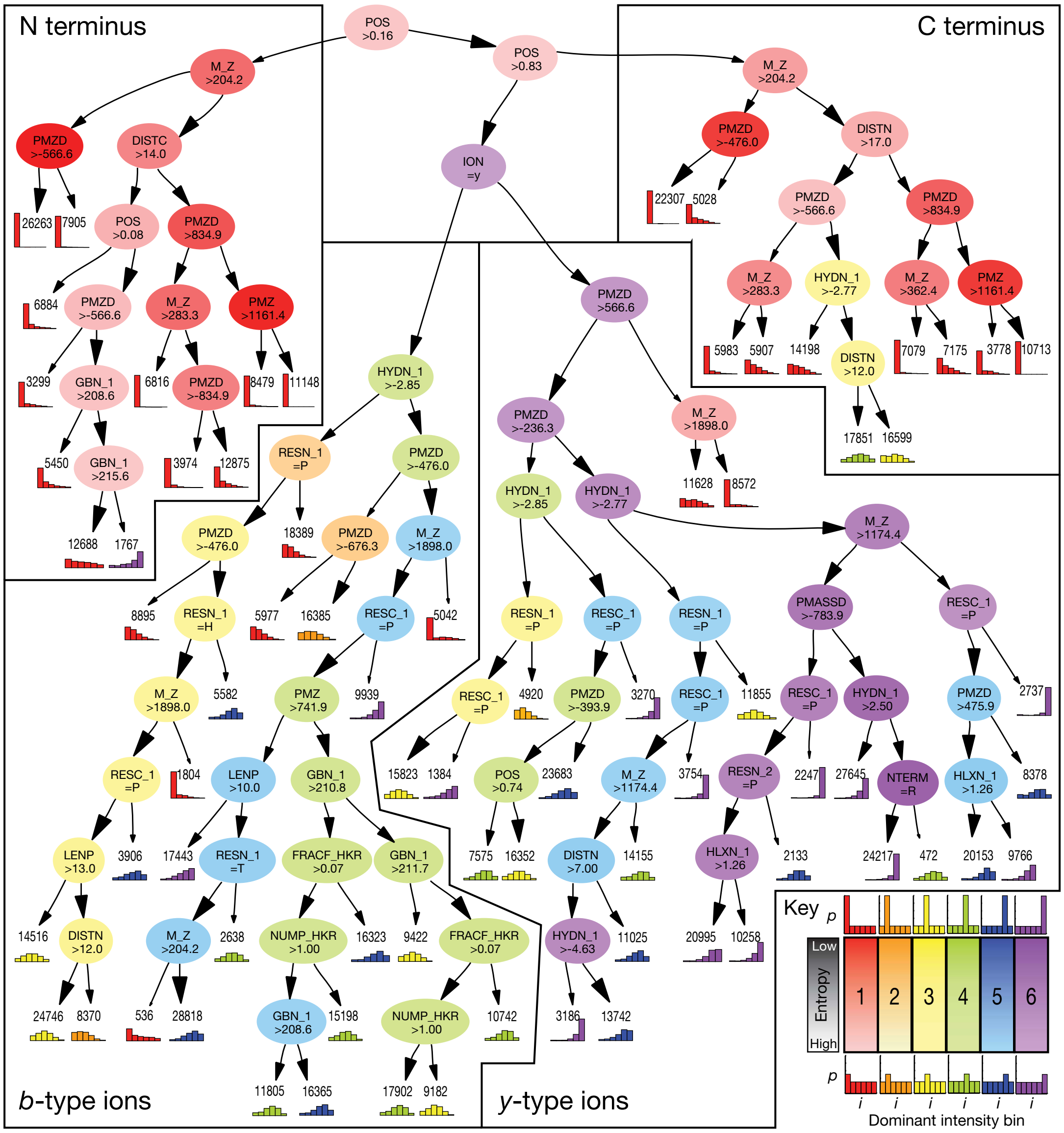

Figure 1 Probabilistic match decision tree automatically learned from fragment intensity data. Text within internal nodes indicates the attribute that provides the greatest reduction in intensity distribution entropy for fragments reaching that point in the tree. Attributes and their abbreviations are described in **Table 1**. Internal nodes (ellipses) or leaf nodes (bar graphs) each have a corresponding probability distribution of intensities. The six colors shown in the key correspond to the six intensity bins (see Methods). Saturated internal node colors indicate probability distributions that strongly favor a single intensity bin (low entropy); lower saturation indicates a weaker tendency to favor a particular intensity bin (high entropy). The relationship between color, saturation and intensity is depicted by the thumbnail graphs of probability (*p*) versus observed intensity (*i*) above (low entropy) and below (high entropy). Arrowhead area is proportional to the fraction of fragments emanating from the source node. Arrows pointing to the right indicate fragments satisfying the condition indicated by the source node; arrows pointing left indicate fragments that do not. The tree is used to assign each individual input fragment to a representative leaf node. Numbers within the leaf nodes indicate the number of training set fragments used to generate the intensity probability distribution represented by the node. This tree was generated from fragment ions of the top SEQUEST-ranked high-confidence peptide-spectral match. Numerical values for attributes related to *m/z*, gas phase basicity, helicity and hydrophobicity were calculated using values found in **Supplementary Table 2** online.

scores (*LOD*, Δ*LOD*), outperforms *Disc*, which contains no *LOD*-based scores. These findings show that *LOD* contributes predictive power not accessible to SEQUEST, providing greater confidence in PSM assignments.

Performance of published SEQUEST criteria (listed in **Table 2**) is also presented in **Figure 3**. The most precise scoring criteria are those used in the recent analysis of the *Plasmodium falciparum* proteome[16]. Applied to our data set of doubly charged, fully tryptic peptides, these criteria yield a precision rate of 99.9%, although this precision rate may not apply to other classes of peptides identified in that study (*e.g.*, partially tryptic or triply charged). Despite this high precision rate, these criteria selected only 65.4% of the estimated correct PSMs in our test set. In comparison, at a threshold yielding the same precision, *CompLOD* has a sensitivity of 81.1%, yielding 23.6% more correct identifications than these criteria and 13.4% more than the *Disc* score. The criteria with the greatest

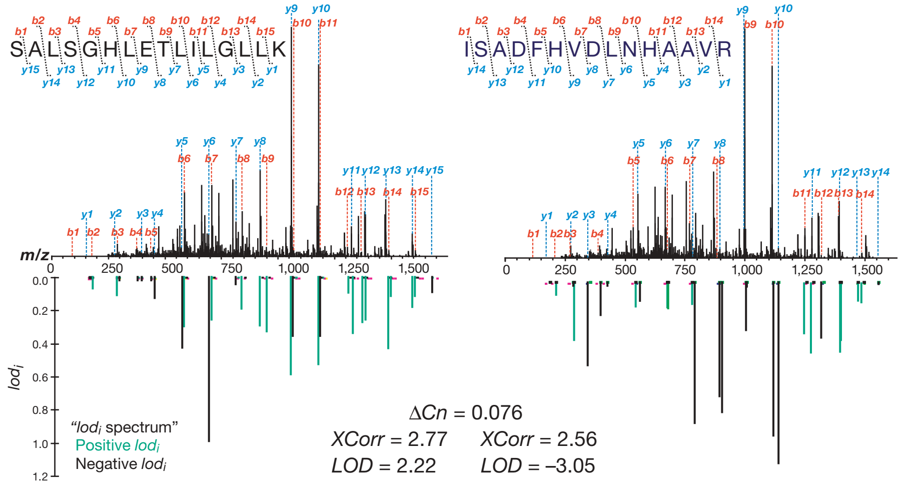

Figure 2 Example of lod_i spectra. Each predicted fragment *m/z* is assigned a lod_i score, plotted on the lower axis below the MS/MS spectrum. Positive values (green) indicate that the fragment is more likely to be derived from a correctly matched PSM than a mismatched PSM. Negative values (black) indicate that the fragment is more likely to be derived from a mismatched PSM than a correctly matched PSM. SEQUEST returned the two peptides shown as the best matches, but the *XCorr* scores were too similar to allow a confident call. By contrast, the *LOD* score strongly indicates that SALSGHLETLILGLLK gave rise to the observed spectrum. Tree navigation and LOD score calculation for the *y10* ion from the SALSGHLETLILGLLK spectrum is shown in **Supplementary Figure 5** online. Similar plots may be obtained for individual spectra using *SILVER*, a web application at http://llama.med.harvard.edu/Software.html/ (F.D.G., J.E.E., S.P.G. and F.P.R., unpublished software).

sensitivity[7] has a precision of 95.9%, identifying 83.5% of the total estimated correct PSMs. *CompLOD* has a precision rate of 99.8% at the same sensitivity. This corresponds to error rate (1 – precision) reductions of 95% and 83% relative to the Peng *et al.*[7] criteria and the *Disc* score, respectively.

This algorithm may be used in conjunction with peptide identification software other than SEQUEST. When this same data set is searched with the Mascot algorithm[17], we find that a composite score incorporating *LOD* decreases error and false negative rates by as much as 96% and 67%, respectively, relative to the Mascot score, and by 83% and 50%, respectively, relative to a combined Mascot score (*AvgMascot*) (**Supplementary Fig. 6** online).

In the absence of a gold standard set of PSMs, we used a collection of PSMs that scored well by SEQUEST. To reduce future reliance on SEQUEST, we might follow an iterative process, in which PSMs with high composite scores are used to retrain the decision tree.

Table 2 Table of scores and their descriptions

Score	Description
CompLOD	Composite score: average of standardized scores: *LOD*, Δ*LOD*, *XCorr*, Δ*Cn*
Disc	Discriminant score[15]
LOD	Log(p(intensity|match)/p(intensity|mismatch))
Δ*Cn*	Normalized difference between first- and second-ranked *XCorr* scores
XCorr	SEQUEST cross-correlation score
Δ*LOD*	Difference between *LOD* scores of first and second peptides as ranked by *XCorr*
A	Published criteria[16]: *XCorr* ≥ 2.5, Δ*Cn* ≥ 0.08
B	Published criteria[21]: *XCorr* ≥ 2.2, Δ*Cn* ≥ 0.10
C	Training set selection criteria: *XCorr* ≥ 2.0, Δ*Cn* ≥ 0.10
D	Published criteria[7]: *XCorr* ≥ 1.5, Δ*Cn* ≥ 0.08

Although the focus above has been on peptide identification, the goal of most proteomics experiments is protein identification. Therefore, we compared the number of proteins identified by each score considered in **Figure 3** and **Supplementary Figure 6** online at the precision levels shown in **Figure 3c** and **Supplementary Figure 6d** online. As expected, higher precision on the peptide level translates to higher protein precision (**Fig. 3b,c** and **Supplementary Fig. 6c,d** online). However, this correlation is not perfect: for the thresholds shown, the protein identification error rate is two to three times greater than the peptide identification error rate. This is because the set of correctly matched peptides maps to fewer proteins (most incorrectly identified proteins are matched only to one peptide)[7]. Indeed, the number of peptides identified for a given protein is a commonly used heuristic measure of confidence in protein identification, and proteins with only a single identified peptide are often discarded. Scores incorporating *LOD* measurements may allow more peptides to be identified for a given protein, or allow high-confidence identification of a protein based on only one matched peptide.

All data presented thus far were acquired in-house on ion-trap tandem mass spectrometers. To determine whether our decision trees are applicable to spectra acquired elsewhere on similar mass spectrometers, we assigned *LOD* scores to another SEQUEST-searched data set[15]. Redundancy filtering applied to this data set yielded 287 spectra. Using parameters measured from our test data set, we assigned *CompLOD* scores to these 287 spectra (Methods). The *Disc* score yielded 49 correct matches before its first false positive (correctness is assumed if the matched peptide belongs to one of the 18 proteins used to generate the spectra). By comparison, the *CompLOD* score matched 68 correctly before its first false positive. This observation further validates our method as one that enhances both sensitivity and precision of PSM selection, and shows that observed fragment ion intensities are not completely instrument dependent. Thus, a relatively unoptimized composite score incorporating *LOD* (*CompLOD*) surpasses a composite score including only SEQUEST-derived scores (*Disc*).

In this work, we present four main results: (i) fundamental ion-trap fragmentation phenomena can be learned from a large collection of high-confidence PSMs; (ii) relationships between peptide and fragment properties and peak intensities in MS/MS spectra can be modeled probabilistically; (iii) combining fragment intensity predictions with existing spectral interpretation methods improves the likelihood of correctly identifying which peptide gave rise to a candidate MS/MS spectrum; and (iv) making use of a decoy proteome approach permits evaluation of sensitivity and precision for alternative scoring strategies without requiring a manually curated test data set. The methods described here both increase the number of spectra for which a confident match can be made and reduce the false positive rate. Although we have applied our approach only to MS/MS spectra acquired with ion-trap mass spectrometers, we expect this approach will apply to other instrument types. However, instrument type–specific training sets may be required.

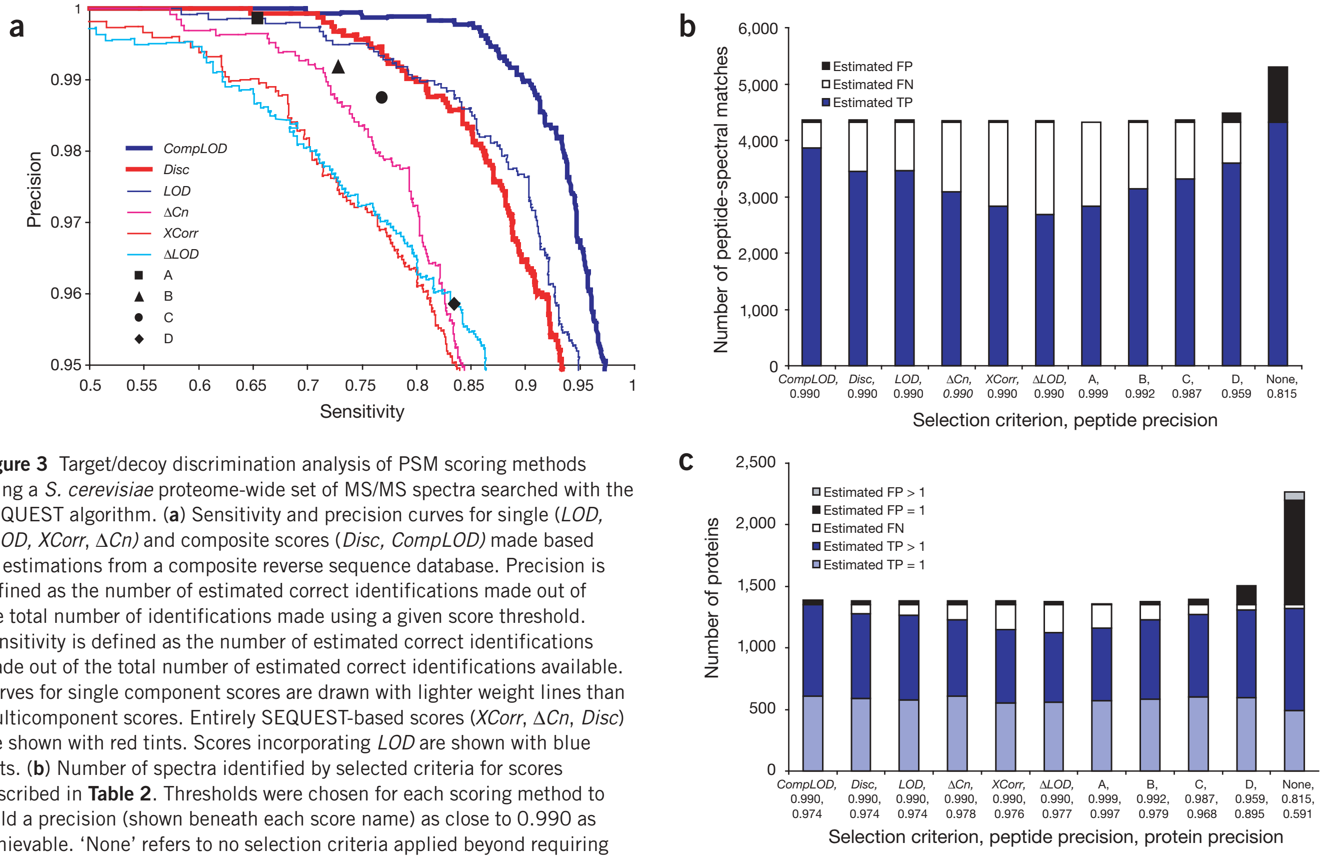

Figure 3 Target/decoy discrimination analysis of PSM scoring methods using a *S. cerevisiae* proteome-wide set of MS/MS spectra searched with the SEQUEST algorithm. (**a**) Sensitivity and precision curves for single (*LOD, ΔLOD, XCorr*, *ΔCn)* and composite scores (*Disc, CompLOD)* made based on estimations from a composite reverse sequence database. Precision is defined as the number of estimated correct identifications made out of the total number of identifications made using a given score threshold. Sensitivity is defined as the number of estimated correct identifications made out of the total number of estimated correct identifications available. Curves for single component scores are drawn with lighter weight lines than multicomponent scores. Entirely SEQUEST-based scores (*XCorr*, *ΔCn, Disc*) are shown with red tints. Scores incorporating *LOD* are shown with blue tints. (**b**) Number of spectra identified by selected criteria for scores described in **Table 2**. Thresholds were chosen for each scoring method to yield a precision (shown beneath each score name) as close to 0.990 as achievable. 'None' refers to no selection criteria applied beyond requiring fully tryptic peptides. Numbers of true and false positives were estimated using a decoy proteome approach as described (**Supplementary Methods** online). The number of correct peptides missed by the score threshold (false negative) was estimated by subtracting the number of TP for a given score from the number of estimated correct TP (**Supplementary Methods** online). FP, false positive; TP, true positive; FN, false negative. (**c**) Number of proteins identified by selected criteria cutoffs in **b**, and the corresponding protein precision (number of correct protein identifications/number of protein identifications). Precisions of peptide and protein identifications are listed below score names. "=1" indicates the class of proteins identified by one peptide; ">1" indicates proteins identified by multiple peptides.

METHODS

Mass spectrometry. Human, mouse, yeast and other protein sources were analyzed over a two-year period through ongoing research within the Gygi Lab and the Taplin Mass Spectrometry Facility. MS/MS spectra were acquired by nanoscale, microcapillary, liquid chromatography–MS/MS as described[18] on either LCQ DECA or DECA XP mass spectrometers (ThermoElectron). Spectra were searched with the SEQUEST[8] or Mascot[17] algorithms (as indicated) against appropriate sequence databases without enzymatic restriction unless otherwise noted.

Training set construction. Doubly charged, top-ranked (by SEQUEST), fully tryptic PSMs receiving *XCorr* scores ≥2.0 and *ΔCn* scores ≥0.10 were selected as high-confidence matches and filtered to remove redundant PSMs (those with identical matching peptides). For each of these top-ranked (matched) as well as second-ranked (mismatched) PSMs, 63 peptide and fragment attributes were recorded. Observed intensities were assigned to predicted fragment ions if the corresponding observed fragment ion's *m/z* was within 1.25 amu of the predicted fragment ion. We used a window of 1.25 amu to allow selection of longer fragment ions that may exceed the commonly applied 1.00 amu window as a result of expected shifts in isotopic distributions in these larger ions. Numerical data were assigned to one of six (for intensity) or a maximum of 20 bins (for other attributes). All unobserved data or data that could not be assigned to a specific bin were assigned to an additional 'undefined' bin. Bin ranges were chosen to equalize the number of fragments assigned to each bin. Fragment intensities were normalized (I_{norm}) so that $I_{norm} = \ln({}^{I_{raw}}\!/\!\sum_i I_i)$ (I_{raw}, raw fragment intensity; I_i, the i^{th} raw fragment intensity). Intensity bin ranges are as follows (normalized intensity units): bin 1, (unobserved); bin 2, [–∞, –6.425); bin 3, [–6.425, –5.600); bin 4, [–5.600, –4.887); bin 5, [–4.887, –4.054); bin 6, [–4.054, 0.000]. All nonnumeric data (*e.g.*, residue identity) were assigned their own discrete bins.

Intensity modeling. Attribute values were calculated for all 783,994 singly charged *b*- and *y*-type fragment ions predicted from the 27,266 peptides in the training set (**Supplementary Fig. 2** online). Shannon entropy was measured for the pre- (parent) and post- (child) split fragment intensities,

$$H(I) = -\sum_j p(I_j)\ln p(I_j)$$

where *H* is entropy, *j* is a given intensity bin index and I_j is the intensity bin *j*. The change in entropy was calculated as,

$$\Delta H = \frac{N}{M}\left(H_{parent} - p(right\ child)H_{right\ child} - p(left\ child)H_{left\ child}\right)$$

where Δ*H* is the change in entropy, *N* is the number of data points at the parent node, *M* is the number of data points at the root node, *H* is the measured entropy at a given node, and *p*(*left child*) and *p*(*right child*) are the proportions of fragments in the parent node assigned to each child. The attribute and bin yielding the greatest reduction in entropy were selected as the next node for the growing decision tree. Branches were terminated when the creation of additional child nodes did not reduce entropy sufficiently to improve the model. According to the Bayesian information criterion[10], entropy reduction (Δ*H*) must be greater than the threshold ln(*M*)/2*N* with *M* and *N* defined as above. Probabilistic decision tree construction took ~2 h on a 2.2 GHz Advanced MicroDevices PC.

Test data set. 81,206 MS/MS spectra previously obtained from an analysis of the *Saccharomyces cerevisiae* proteome[7] were searched with SEQUEST and Mascot algorithms against a sequence database containing translated yeast open reading frames in both forward (correct, 'target') and reverse (incorrect, 'decoy') orientations as described[7]. We estimated precision (selected correct PSMs/all selected PSMs) and sensitivity (selected correct PSMs/all correct PSMs) for several scoring methods (**Supplementary Methods** online). Doubly charged, top-ranked (by SEQUEST or Mascot), fully tryptic, nonredundant PSMs were selected as test data.

Composite score. The *CompLOD* and *CompLODm* scores combining multiple score types were generated as follows: each score's mean (μ) and standard deviation (σ) were measured from the top-ranked (by either SEQUEST or Mascot), doubly charged, fully tryptic, nonredundant peptides used as the yeast proteome test set. Individual scores (x_i) were standardized and averaged to yield the composite score as follows:

$$Comp = \frac{1}{N_x}\sum_{i}^{N_x}\frac{x_i - \mu_i}{\sigma_i}$$

where i is a particular score type (*e.g.*, *XCorr*, *LOD*) and N_x is the number of score types.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank the past and present members of the Gygi Lab and Taplin Biological Mass Spectrometry Facility for generating the spectra used in this study. We gratefully acknowledge John Cottrell of Matrix Science, for examining our test data set with the Mascot algorithm. Particular thanks to S. Gerber, J. Jebanathirajah and H. Steen for insightful comments during manuscript preparation. This work was supported in part by National Institutes of Health (NIH) HG00041 (S.P.G.), NIH National Research Service Award 5T32CA86878 from the National Cancer Institute (J.E.E.), by an institutional grant from the Howard Hughes Medical Institute (F.P.R., F.D.G.), and by National Research Service Award fellowship from NIH/National Human Genome Research Institute (O.D.K.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 11 August; accepted 31 October 2003
Published online at http://www.nature.com/naturebiotechnology/

1. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
2. Mann, M., Hendrickson, R.C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, 437–473 (2001).
3. Aebersold, R. & Goodlett, D.R. Mass spectrometry in proteomics. *Chem. Rev.* **101**, 269–295 (2001).
4. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
5. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193–197 (2003).
6. Gay, S., Binz, P.A., Hochstrasser, D.F. & Appel, R.D. Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra. *Proteomics* **2**, 1374–1391 (2002).
7. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. & Gygi, S.P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50 (2003).
8. Eng, J., McCormack, A. & Yates, J.R. 3rd. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
9. Jensen, F.V. *Bayesian Networks and Decision Graphs* (Springer, New York, 2001).
10. King, O.D., Foulger, R.E., Dwight, S.S., White, J.V. & Roth, F.P. Predicting gene function from patterns of annotation. *Genome Res.* **13**, 896–904 (2003).
11. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech J.* **27**, 379–423,623–656 (1948).
12. Papayannopoulos, I.A. The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrom. Rev.* **14**, 4973 (1995).
13. Breci, L.A., Tabb, D.L., Yates, J.R. 3rd & Wysocki, V.H. Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal. Chem.* **75**, 1963–1971 (2003).
14. Tabb, D.L. *et al.* Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* **75**, 1155–1163 (2003).
15. 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.* **74**, 5383–5392 (2002).
16. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526 (2002).
17. Perkins, D., Pappin, D., Creasy, D. & Cottrell, J. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
18. Peng, J. & Gygi, S.P. Proteomics: the move to mixtures. *J. Mass Spectrom.* **36**, 1083–1091 (2001).
19. Harrison, A.G. The gas-phase basicities and proton affinities of amino acids and peptides. *Mass Spectrom. Rev.* **16**, 201–217 (1997).
20. Deber, C.M. *et al.* TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci.* **10**, 212–219 (2001).
21. Washburn, M., Wolters, D. & Yates, J.R. 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).