# Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry

Joshua E Elias[1] & Steven P Gygi[1,2]

**Liquid chromatography and tandem mass spectrometry (LC-MS/MS) has become the preferred method for conducting large-scale surveys of proteomes. Automated interpretation of tandem mass spectrometry (MS/MS) spectra can be problematic, however, for a variety of reasons. As most sequence search engines return results even for 'unmatchable' spectra, proteome researchers must devise ways to distinguish correct from incorrect peptide identifications. The target-decoy search strategy represents a straightforward and effective way to manage this effort. Despite the apparent simplicity of this method, some controversy surrounds its successful application. Here we clarify our preferred methodology by addressing four issues based on observed decoy hit frequencies: (i) the major assumptions made with this database search strategy are reasonable; (ii) concatenated target-decoy database searches are preferable to separate target and decoy database searches; (iii) the theoretical error associated with target-decoy false positive (FP) rate measurements can be estimated; and (iv) alternate methods for constructing decoy databases are similarly effective once certain considerations are taken into account.**

One of the most daunting tasks researchers face when conducting mass spectrometry–based proteomics investigations is the correct selection of the 10–50% of spectrum assignments generated in LC-MS/MS experiments that are actually correct[1]. Unlike Edman degradation and automated DNA sequencers that build polypeptide and DNA sequences *de novo* by sequential identification of individual monomers, mass spectrometry–based identifications are usually inferred from the quality of the match between observed and predicted sequence-specific patterns. Although methods for generating expected MS/MS spectra and comparing them to observed spectra are continually evolving, matches between observed and expected spectra are rarely ideal. Furthermore, peptide MS/MS spectra interpretation can be problematic when low-intensity nonpeptide species are selected for fragmentation, the peptides being examined are not anticipated in the sequence database being queried, or the MS/MS spectra are not of sufficient quality for definitive interpretation. For these reasons, some degree of ambiguity is usually associated with each peptide identification. Several options have been described to reduce this ambiguity, including case-by-case inspection[2] and applying scoring filters based on how the filters performed on an often smaller and unrelated training data set[3]. In our experience, the target-decoy search strategy provides a substantial improvement over these options—both in throughput and accuracy—and is the best available option for analyzing large-scale studies.

Rather than deciding exactly which peptide-spectral matches (PSMs) are correct or incorrect, the composite target-decoy database evaluates FP rates in large PSM populations. It permits estimation of the likelihood that a PSM is correct given that it came from a collection of PSMs with a measured FP rate. This is not to suggest that the search strategy removes all false identifications. Instead, the target-decoy approach allows the estimation of how many FP are associated with an entire data set. Furthermore, this method can suggest PSM attributes (for example, peptide length, elution time, charge, algorithm-assigned score) that distinguish correct identifications.

The target-decoy strategy is simple to implement in principle (**Supplementary Fig. 1** online): A composite database is created by first obtaining a 'target' protein sequence database appropriate to the protein mixture to be analyzed. Next, a 'decoy' database is created to preserve the general composition of the target database while minimizing the number of peptide sequences in common between the target and decoy. This is most

[1]Department of Cell Biology and [2]Taplin Biological Mass Spectrometry Facility, 240 Longwood Avenue, Harvard Medical School, Boston, Massachusetts 02115, USA. Correspondence should be addressed to S.P.G. (steven_gygi@hms.harvard.edu).

PUBLISHED ONLINE 27 FEBRUARY 2007; DOI:10.1038/NMETH1019

**Table 1** | Measurements derived from decoy database search results

| Measurement | Formulation | Description of estimate |
|---|---|---|
| False positive (FP) | 2 × passing decoy assignments | Number of incorrect assignments above score threshold |
| True positive (TP) | Total passing assignments – number of FPs | Number of correct assignments above score threshold |
| Total correct (TC) | Maximum TPs for all evaluated score criteria combinations | Number of total correct assignments in the data set |
| Total incorrect (TI) | Total assignments – TC | Number of total incorrect assignments in the data set |
| False negative (FN) | TC – TP | Number of correct assignments falling below score threshold |
| True negative (TN) | TI – FN | Number of incorrect assignments falling below score threshold |
| Precision | TP / (TP + FP) | Fraction of correct assignments above score threshold |
| FP rate | FP / (TP + FP) or 1 – precision | Fraction of incorrect assignments above score threshold |
| Sensitivity | TP / TC | Fraction of all correct assignments above score threshold |
| Specificity | TN / (TN + FP) | Fraction of all incorrect assignments below score threshold |
| Accuracy | (TP + TN) / total assignments | Fraction of all assignments correctly classified by score threshold |



simply done by reversing the target protein sequences[4–6], although decoy databases can also be created by stochastic means, such as Markov chain modeling[7]. Decoy sequences must be clearly labeled so they can be distinguished from target sequences in the search results. Appending the decoy database to the target database creates a composite database twice the size of the original. MS/MS spectra are searched against this single composite database. Assuming no correct peptides are found in both target and decoy portions, and that incorrect assignments from target or decoy sequences are equally likely, one can estimate the total number of FPs that meet specific selection criteria by doubling the number of selected decoy hits. This represents the number of obvious incorrect decoy hits, combined with the hidden incorrect target hits. With FP estimations, it is possible to derive other measurements that help evaluate and compare scoring methods and data sets (**Table 1**). All measurements are reported in the context of the entire selected nonredundant data set. Nonredundant data sets can be preferable since they remove the potential of a few abundant, frequently observed peptides (correctly identified or not) from skewing the final FP estimations.

Here we provide evidence that (i) the two assumptions mentioned above are reasonable; (ii) searches should be performed against a single concatenated database rather than parallel searches against target and decoy databases; (iii) the theoretical error associated with FP measurements can be estimated; and (iv) alternate decoy database construction methods are effective, once certain considerations are taken into account. We then suggest ways in which decoy hits can be used to tailor selection criteria to individual data sets.

## Assumption 1: target and decoy databases do not overlap

If we are to trust that decoy hits are incorrect, we must be confident that they are not present in the (ideally) comprehensive target database. This was investigated *in silico* by generating a list of peptides that result from a proteolytic (for example, tryptic) digest of the target proteome, and comparing this list to that from a similarly digested decoy sequence database. As expected, very short peptides were found in both target and decoy databases. The proportion of sequences in common, however, decreased precipitously with increasing peptide length such that practically no (0.02%) peptides with lengths greater than eight amino acids were in common between target and decoy databases. For all databases we have examined, the average tryptic peptide length is greater than nine amino acids, suggesting this first assumption is reasonable (**Fig. 1**). Furthermore, short peptides (less than eight amino acids) are usually underrepresented in final data sets relative to the expected peptide distribution, because of their propensity to form singly-charged ions, their often polar nature (preventing retention under typical high-performance liquid chromatography conditions) and relatively few fragment ions formed under MS/MS conditions (needed by database search engines). Consequently, the practical effect of rare overlapping peptides should be appreciably less than that suggested by predicted peptide distributions.

Some may be concerned that there is a chance that identified decoy peptides not present in the target database are correct, owing to situations such as mutations, unknown alternative splice variants or contaminant proteins. For decoy peptides to correctly match input MS/MS spectra, however, their sequences would need to have arisen by the practically random process that gave rise to the decoy

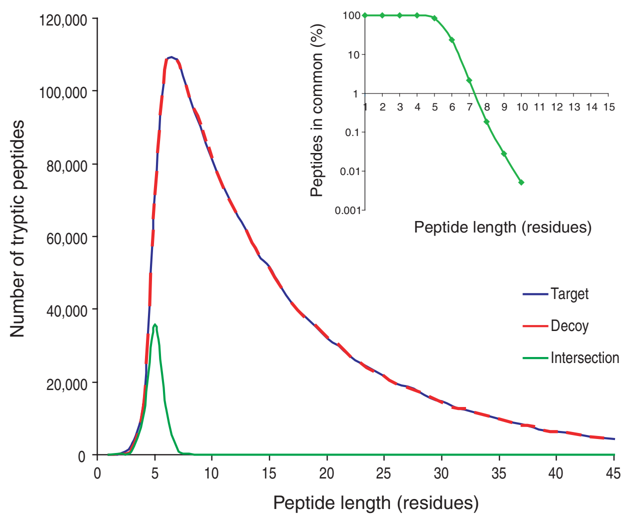



**Figure 1** | Overlap between target (forward) and decoy (reversed) sequences is negligible. Human protein sequences within the minimally redundant International Protein Index sequence database[21] were digested *in silico* with trypsin (maximum two missed cleavage sites, maximum peptide length = 45; target). Tryptic peptides were similarly generated from the reversed protein sequences from this database (decoy). After converting isoleucines to leucines, the number of peptide sequences in common between the two databases was determined (intersection). Practically no peptides greater than 8 amino acids in length were found in both forward and reversed databases. Inset, percentage of peptides in common between target and decoy sequences decreases with increasing peptide length.

database. Although this phenomenon is difficult to model, it clearly should be an exceedingly rare event, insignificant in relation to the thousands or tens of thousands of identifications frequently used to estimate observed FP rates.

### Assumption 2: target and decoy false positives are equally likely

With the composite target-decoy database strategy, the total number of FPs that pass a given set of criteria are estimated by doubling the total number of decoy hits. This is appropriate only if there is an equal likelihood of selecting an incorrect peptide match from target and decoy portions of the composite database. The validity of this assumption can be tested in two ways. First, the search algorithm must be presented with equal numbers of target and decoy peptides. Second, the number of necessarily incorrect peptide hits should be equally distributed between target and decoy hits. These assertions are demonstrated with *in silico* digests (**Fig. 1**), and with a LC-MS/MS data set derived from a complex peptide mixture.

MS/MS search engines like SEQUEST[3] and Mascot[8] evaluate all peptide sequences with masses within a specified range as potential matches to an input spectrum. To demonstrate that the number of considered peptides is equal for target and decoy sequences, we compared *in silico*–digested target and decoy databases within varying mass tolerances. The distributions of considered peptides were practically the same between target and decoy peptides, regardless of mass tolerance (**Fig. 2a**,**b**). Comparing these curves indicated substantial correspondence between target- and decoy-derived peptides, with slopes only slightly deviating from unity (**Fig. 2c**). We observed this correspondence in actual search results as well (**Supplementary Fig. 2** online). It is notable, however, that the agreement in peptide number decreased somewhat with higher mass accuracy, suggesting that the most accurate FP estimations may be derived from lower-tolerance searches. This is expected considering that with higher mass accuracy, there are fewer combinations of amino acids that can produce an observed mass. Therefore, with relatively few candidate sequences, the degree of correlation between target and decoy hits tends to be noisier.

Although many peptides are usually considered as potential matches to input MS/MS spectra, typically the one receiving the highest algorithm-assigned score is selected as the most likely match.



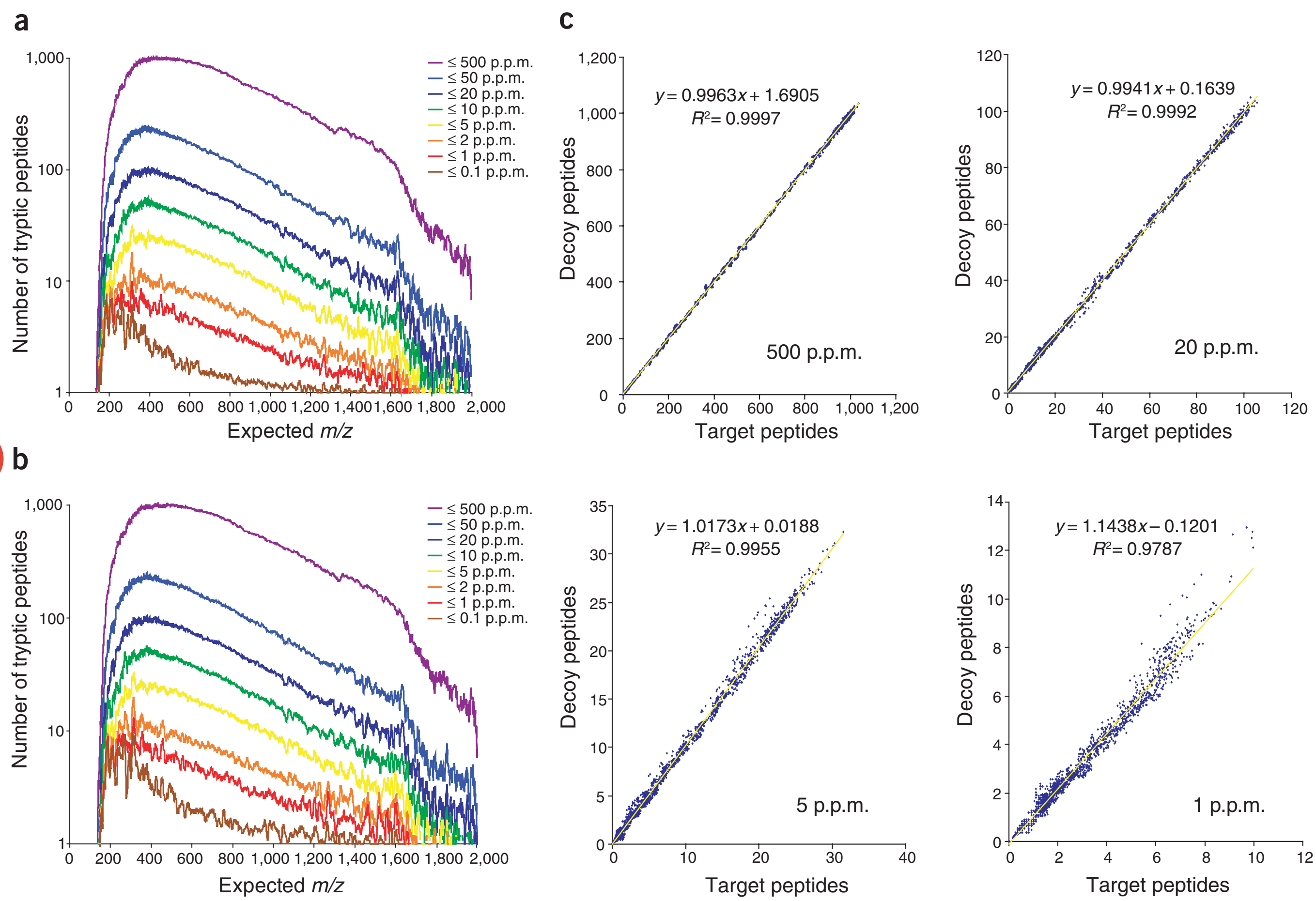



**Figure 2** | The distributions of potential peptide matches is consistent between target and decoy databases at several mass tolerances. (**a**,**b**) Human protein sequences within the minimally redundant International Protein Index sequence database were digested in silico with trypsin (maximum two missed cleavage sites, maximum peptide length = 45). Expected peptide monoisotopic mass-to-charge ratios (*m*/*z*) were calculated and assessed as to how many distinct tryptic peptide sequences (excluding isobaric isoleucine and leucine differences) fall within 500, 50, 20, 10, 5, 2, 1 and 0.1 p.p.m. mass units of each peptide. Moving averages (±5 mass units) were plotted for each p.p.m. range. Target database (**a**). Decoy database (reversed; **b**). (**c**) Direct comparisons of the indicated data plotted in **a** and **b**. The least-squares best-fit line has a slope that deviates only slightly from unity, with strong correspondence to the data ($R^2 > 0.97$), although the similarity between target and decoy slightly decreases with increasing mass accuracy.

As the majority of correct peptide identifications achieve the top rank, lower-ranked peptides are largely incorrect. If target and decoy peptides are equally likely to be incorrectly selected as matches, these lower-ranked peptides should be evenly distributed between the two database portions. We illustrate this phenomenon in **Figure 3a**, indicating that top-ranked peptides showed a strong bias toward target database hits, unlike lower-ranked matches. We extended this idea by modifying MS/MS spectra to prevent any correct identifications from being made (**Fig. 3b**). In this situation, target and decoy peptides were selected with near equal frequency regardless of rank.

### Concatenated database searches are preferable to separate searches

Several research groups have adopted a target-decoy strategy similar to its first description[4] in which decoy sequences are searched separately in order to assess score distributions obtained from apparently random matches[9–12]. Ostensibly, this strategy seems reasonable: if decoy sequences are incorrect, they should behave identically whether they are considered with target sequences or by themselves. However, when MS/MS spectra are searched in this way, target and decoy sequences cannot compete for the top-ranked score in a single search. Without competition, decoy sequences that partially match to high-quality MS/MS spectra may often receive elevated scores relative to other top-ranked hits from the search, suggesting high filtering thresholds need to be applied (**Supplementary Fig. 3** online).

Another option is to search MS/MS spectra once against a single database consisting of target and decoy sequences. Besides decreasing search times by 20–30% relative to two separate searches (data not shown), target and decoy sequences directly compete for the algorithm-assigned top rank for a given MS/MS spectrum. Under these circumstances, target and decoy sequences can usually be easily distinguished despite a relatively high decoy score (**Supplementary Fig. 3**). Because relatively high-scoring decoy sequences rarely outscore correct identifications in the composite database scenario[1], researchers will not be misled to set inappropriately high scoring criteria.

Notably, separate searching greatly obstructs the ability to estimate low-scoring correct identifications in the presence of high-scoring incorrect identifications (**Fig. 4**). The separate search method forces one to assume that all peptide assignments are incorrect below the score at which decoy hits outnumber target hits, leading to an overestimated FP rate[13] (**Fig. 4e**). This is not the case with target-decoy searching, permitting the estimation of correct identifications obscured by high-scoring decoy identifications (**Fig. 4d**).

It should be noted that searches against a composite target-decoy database yield relative scores (for example, delta corretation ($\Delta Cn$) for SEQUEST or the homology threshold for Mascot) that can differ from searches against the separate target or decoy databases. Although these scores can sometimes vary widely, we found they did not substantially change for more than half of all considered peptides. Where we observed shifts between composite and separate searches, they were proportionally similar between target and decoy databases. These observations were fairly consistent between Mascot and SEQUEST (**Supplementary Fig. 4** online). Therefore we feel that shifts in relative score shifts caused by composite database searching are a reasonable cost considering the overall benefit of improved estimation of true and false identifications. It is fair to say, however, that derived filtering criteria involving these relative scores should not be applied to data searched against target databases alone.

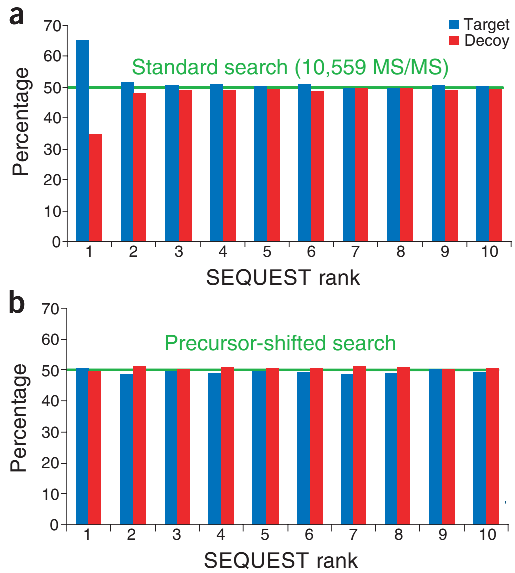



**Figure 3** | Incorrect identifications are equally selected from target and decoy (reversed) sequences. (**a**) More than 10,000 MS/MS spectra derived from a complex mixture of trypsin-digested Jurkat lysate were searched with the SEQUEST algorithm against a composite target-decoy sequence database with a mass tolerance of 50 p.p.m. The number of target and decoy hits reaching each rank was counted. A roughly one-to-one correspondence between target and decoy hits was observed for ranks two through ten, which are overwhelmingly incorrect identifications. (**b**) The spectra described in **a** were altered by shifting their precursor ion $m/z$ values by 10 Da, rendering them impossible to be correctly matched by SEQUEST. Consequently, all ranks were incorrect. These results are similar to analyses performed on unrelated data using different search conditions[19,20].

### Estimating theoretical error of target-decoy false positive rates

One criticism of the target-decoy approach is that one can never know exactly which or how many selected PSMs are incorrect. Although doubling the number of decoy hits provides a reasonable estimation under most conditions, one might expect these estimations substantially deviate from the actual number of FPs when the number of returned hits is very small or the number of returned decoy hits is very large. To gain a better understanding of this issue, we simulated many target-decoy hits *in silico* assuming set numbers of correct identifications and total returned hits, and recorded the variation in the difference between calculated and expected precision rates. Based on these findings, it was possible to place confidence intervals on target-decoy estimations given the number of total hits returned and the estimated precision rate derived from the decoy hits (**Fig. 5a** and **Supplementary Methods** online).

Two important conclusions can be drawn from this computational experiment. First, as expected, the accuracy of the target-decoy approach was diminished greatly with smaller numbers of total considered hits, but accuracy approaches perfection with increased total hits. Second, estimations of FPs were less reliable when data sets contained large proportions of incorrect identifications.

Log-transforming the standard deviation and sample size values allowed straight-line approximations of the curves shown in the graph in **Figure 5a** with slopes that decrease with increased precision (**Fig. 5b** and **Supplementary Table 1** online). By plotting these slopes against ln(1– precision) (**Fig. 5c**), the relationship between the slopes in the plot in **Figure 5b** and precision can be inferred according to equation 1.

$$m = \frac{\ln(1 - \text{precision})}{15} - 0.5 \quad (1)$$

The plot in **Figure 5b** indicates the relationship between the standard deviation (σ ) of the error associated with a given precision measurement and the sample size *(N)* can be approximated by equation 2.

$$\ln(\sigma) = m \ln(N) \quad (2)$$

Therefore, the expected standard deviation of the error associated with an estimated precision rate error can be calculated given the number of total hits considered using equation 3.

$$\sigma = e^{\left(\frac{\ln(1 - \text{precision})}{15} - 0.5\right)\ln(N)} \quad (3)$$

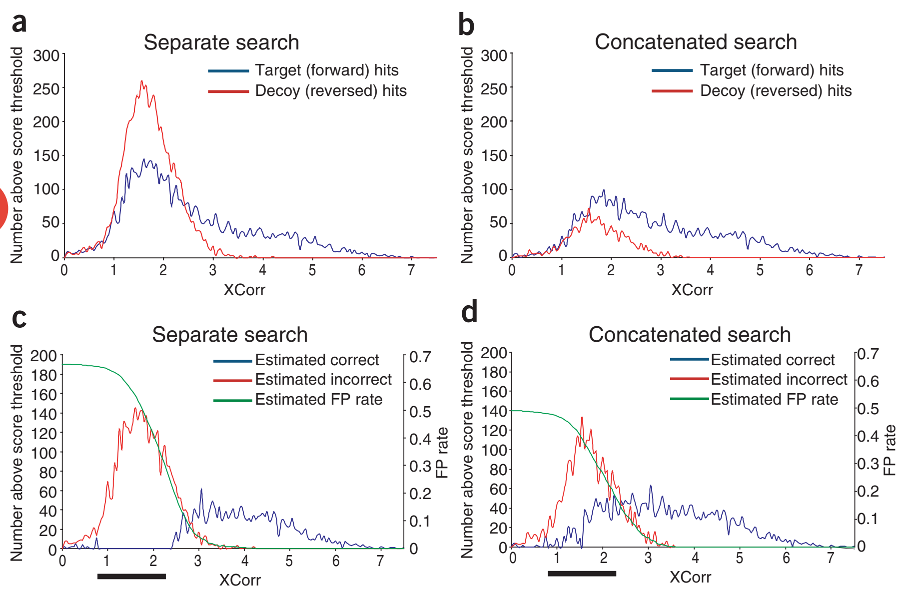
Typical results from searching 10,000 MS/MS spectra may yield 2,500 nonredundant PSMs filtered to have an estimated 1% FP rate based on passing decoy hits. Using equation 3, the standard deviation of the measurement error associated with this actual precision is 0.18%. This suggests the actual FP rate should fall between 0.77% and 1.23% within a 99% confidence interval. Conversely, equation 3 implies that if no filtering were applied to a set of 10,000 spectra and only 100 were estimated to be correct, this observation may have arisen by chance: the standard deviation (0.99%) for this FP rate (99%) and a 99% confidence interval threshold indicate as few as zero and as many as 228 correct identifications could have given this result. Furthermore, assuming there were zero correct identifications in this data set, equation 3 suggests that as many as 129 correct identifications could be estimated by chance. These examples demonstrate how minimum thresholds on estimated correct identifications for sets or subsets of PSMs can be set.

## Alternate decoy database constructions can be similarly effective

Protein sequence reversal is an attractive method for generating decoy sequence databases owing to its simplicity and effectiveness. With little foresight, the reverse transformation preserves amino acid frequencies, protein and (approximate) peptide length distributions as well as approximate mass distributions of theoretical peptides. This transformation of protein sequences, however, is not truly random as some might prefer for an analysis of this kind, nor are the numbers of target and decoy peptides constrained to be equal (**Supplementary Fig. 2**). Consequently, the previously described assumptions, though generally valid, cannot be considered rules. In addition to direct protein reversal, we evaluated one modified sequence reversal method and two stochastic methods that separately address these two issues (**Fig. 6**).

The version of SEQUEST implemented on the Sorcerer platform (SAGE-N research) permits the construction of 'pseudo-reversed' decoy peptide sequences matched to all target sequences (**Fig. 6a**). Thus, by rearranging the amino acids of all target peptides to create decoy peptides, the number of considered target and decoy peptides must be equal. As expected, incorrect identifications were equally distributed between target and decoy peptide sequences in an example data set (**Fig. 6b**).

The same could not be said for the two stochastic methods used for decoy database construction (**Fig. 6b**). One method relied



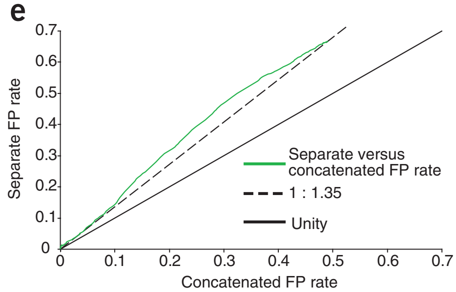




**Figure 4 |** Separate searching overestimates FP rates by underestimating low-scoring correct identifications. (**a**–**d**) More than 6,000 MS/MS spectra derived from a fraction of yeast lysate were searched against separate (**a**,**c**) or concatenated (**b**,**d**) target and decoy sequence databases with the SEQUEST algorithm. Histograms of the number of target or decoy peptides that meet XCorr criteria (bin size = 0.05 units; **a**,**b**). In **c**, estimated correct identifications from separate searches were calculated by subtracting the numbers of decoy peptides from target peptides at given XCorr thresholds. Incorrect identifications were estimated as the minimum of the number of target or decoy peptides returned at given XCorr thresholds. Estimated FP rates (cumulative) were calculated by dividing estimated incorrect peptides by the number of target peptide hits that surpass each XCorr threshold. Notably, this search method cannot estimate correct identifications when decoy hits outnumber target hits (bar spanning 0.8–2.3). In **d**, estimated correct identifications from concatenated searches were calculated by subtracting twice the number of decoy peptides from the total peptides returned at given XCorr thresholds. Incorrect identifications were estimated by doubling the number of decoy peptides returned at given XCorr thresholds. Estimated FP rates (cumulative) were calculated as in **c**. With this method, target and decoy sequences compete, making it possible to estimate the distribution of low-scoring correct identifications (bar spanning 0.8–2.3). (**e**) Direct comparison of FP rates in **c** and **d** indicate that separate database searches can overestimate FP rates by >35% relative to concatenated searches.

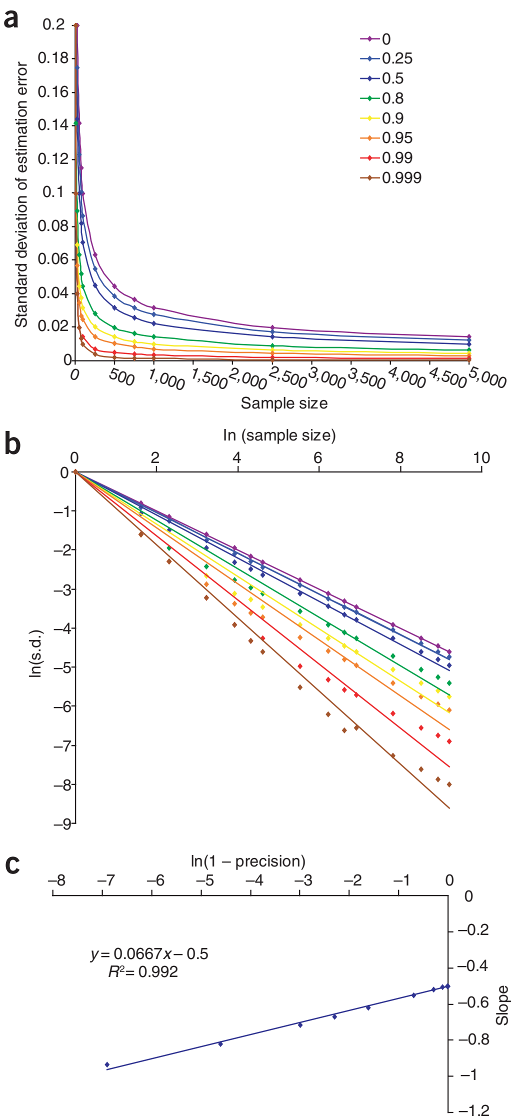



**Figure 5** | Estimating the theoretical error associated with target-decoy estimations. (**a**) Software was written to simulate FP estimations derived from set numbers of correct and incorrect identifications. The program took as input the number of total hits to consider, and what portion are actually correct (that is, precision; **Table 1**). Correct hits were assigned a "target" state. The program randomly assigned each of the remaining incorrect hits a 'target' or 'decoy' state. Once all hits were assigned a state, the precision rate was calculated by doubling the number of decoy hits and dividing this by the total number of hits. This number was subtracted from the predetermined precision rate to give the deviation between the actual and estimated precision rates (estimation error). This process was repeated 100,000 times, creating a distribution of estimation error from which a standard deviation was derived. Such standard deviation measurements were made for many combinations of input total hits and precision. The variation in target-decoy precision rate estimations is dependant on the sample size (number of considered peptide-MS/MS spectrum matches), and the actual precision rate of the data set. Larger standard deviation indicates less reliable precision rate estimations. (**b**) Straight (least-squares) lines fit to the log-transformed data in **a**. (**c**) The slopes of these lines (**Supplementary Table 1**) are related to the underlying precision rate. Combining the trends identified in **Supplementary Table 1** and **c** suggest that the expected standard deviation of a precision rate estimation can be calculated from the precision rate and sample size (equation 3).

on simple amino acid frequencies in the target database to construct the decoy database, constrained only by the target database size and length distribution of its constituent proteins. The other used the same constraints, but relied on a Markov chain model to select which amino acid should follow a particular short amino acid sequence, based on the frequencies found in the target sequence database. Both stochastic methods performed essentially identically to one another in terms of the distribution of target and decoy sequences being incorrectly matched (**Fig. 6b**). Notably, this proportion was not the desired 50%, but was decidedly skewed toward the decoy sequences (63%). Although both random and Markov chain databases were constrained to have similar amino acid compositions as the target database, they clearly produced more peptides than were presented by the target (**Fig. 6c**). We attribute this effect to motifs and domains multiply present in the target database, which reduce the total number of unique sequences considered by the search algorithm. Moreover, even distributions of trypsin cleavage sites (lysine and arginine) may also contribute to this observation by increasing the total number of unique decoy sequences with lengths amenable to identification by mass spectrometry. Sequence reversal reproduces the extent of redundant peptides in the decoy database, whereas stochastic decoy databases do not. This does not negate the utility of stochastic databases, however. By measuring any inherent bias imposed by the particular method of decoy database construction (for example, **Fig. 6b**), it is possible to derive the appropriate factor for estimating FP identifications. Generally, this factor should be equal to the reciprocal of the frequency of unique decoy peptide sequences (that is, 1/(0.63)). Despite their differences, the four decoy databases considered here—protein reversal, peptide pseudo-reversal, random and Markov chain—yielded similar estimations of total correct identifications, and produced similar numbers of correct identifications that can be selected with score criteria designed to yield an estimated 98.0% precision rate (**Fig. 6d**).

## Outlook

For any analytical tool to be truly useful on a large scale there must be a convenient way to assess the validity of its results. This is particularly true for peptide and protein identifications by MS/MS. Typical experiments can yield several thousand identifications, each potentially generating a biological hypothesis worthy of further, in-depth study. It is therefore essential to have a robust method for prioritizing which identifications are most likely to be correct so that investigators do not spend valuable time and resources on misguided follow-up studies. Such prioritization can be done fairly well using scores produced by the search algorithm alone. Scores assigned to peptide hits, however, rarely correspond to their biological importance. Contaminant peptides from keratins, trypsin and certain housekeeping proteins are often among the highest-scoring results, largely owing to their relatively high abundance. At the same time, many correctly identified and biologically interesting peptides may be assigned scores that prevent their unencumbered selection from a sea of incorrect identifications. Notable examples of this are peptides derived from low-abundance transcription factors or peptides that do not fragment well, such as those modified by phosphorylation[14]. As discussed, these low-scoring peptides may be rescued by manually evaluating each identification, either by reanalysis of synthetic peptides, or by expert validation of computer-assigned spectrum interpretations. From a practical point of view, these options are best applied to much smaller systems than the large-scale studies that are the subject of this report.

This report demonstrates that the target-decoy search approach provides a robust and effective way to estimate the number of incorrect identifications that pass any set of filtering criteria. Although this is a crucial measurement for assessing the quality of MS/MS results, it is not in itself a proactive device for establishing a precise, sensitive enrichment of correct peptide identifications from a collection of largely incorrect assignments. Researchers still need to derive score criteria that strike the desired balance between sensitivity and precision in their final data set.

The original publication of the SEQUEST algorithm compared score distributions obtained from a small set of known correct identifications versus known incorrect identifications. From these comparisons, the authors established general guidelines for enriching correct identifications[3]. These guidelines created a precedent for applying discrete score criteria, empirically derived from a small known training data set, to multiple large experimental data sets. As we and others have since shown, static discrete criteria produce variable precision and sensitivity depending on the composition of the searched data[6,15,16]. Probabilistic approaches to assessing PSM validity have advanced our ability to confidently and sensitively select correct identifications[8,16,17], but these methods generally rely on presupposed underlying (and potentially unrelated) probabilistic models[13]. The target-decoy approach, however, makes no modeling assumptions beyond the basic ones validated here.

The approach we used to generate the graph in **Figure 6d** had the goal of producing optimized and tailored discrete score criteria that deliver peptide identifications with a desired FP rate. Similar in principle to a previously described algorithm[6], this method allowed us to evaluate many filtering criteria combinations to arrive at a single combination that yielded a specified minimal precision rate while allowing the greatest number of passing identifications. The agreement between the four decoy database types (**Fig. 6**) indicated that decoy hits can be used to arrive at data set–specific filtering criteria in an unsupervised fashion without overfitting.

This unsupervised filtering application exploits the principle that decoy hits can be used to identify any peptide attribute that provides information on whether a given peptide identification is correct, without prior modeling. These attributes are typically score-based (**Supplementary Fig. 5** online), but can also be spectrum and sequence features that may not be exploited by scoring algorithms, such as charge and the number of internal missed cleavage sites (**Supplementary Fig. 5**). Considering more peptide attributes increases the ability of decoy hits to signify which identifications are most likely to be incorrect, not just how many incorrect identifications are likely to be present. It is tempting to eliminate obviously incorrect decoy hits from a filtered data set. For the reasons described above, however, this may be disadvantageous, as decoy hits can continue to provide information should one choose to revisit a previous analysis.

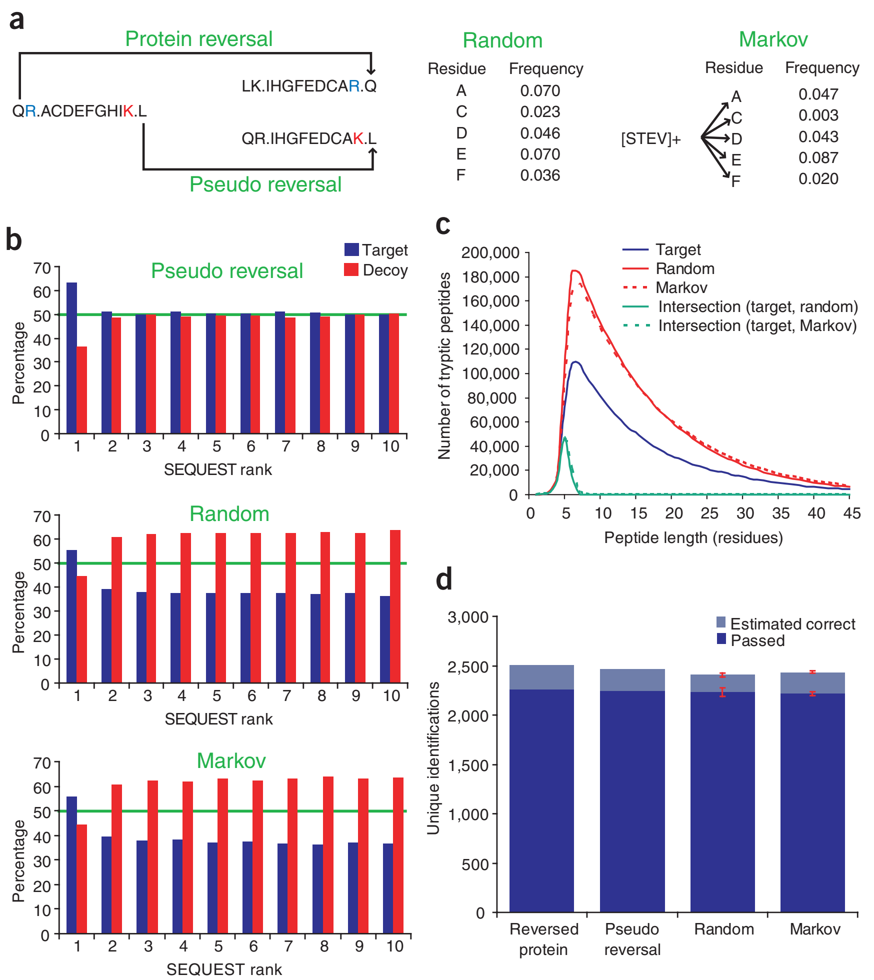



**Figure 6** | Evaluation of alternate decoy databases. (**a**) MS/MS spectra described in **Figure 3** were searched against alternate composite target-decoy databases in which the decoy portion was constructed at search time by peptide pseudo reversal, randomly according to amino acid frequencies in the target database or using a Markov chain model (word length = 5) to preserve a degree of low-order sequence structure in the decoy database. For random and Markov databases, protein lengths were constrained to resemble the distribution found in the target. Comparative depictions of protein reversal, pseudo reversal, random sequence construction and Markov-chain sequence construction. Pseudo-reversal happens at the level of the algorithm; the other methods describe database preparations before searching. (**b**) Frequency of target and decoy identifications for multiple ranks, similar to that in **Figure 3**. (**c**) Length distributions for random- and Markov-created tryptic peptides, similar to that in **Figure 1**. (**d**) Estimated correct identifications after searching with each target-decoy database type. Correct identification estimations and passing score criteria were determined from optimized XCorr and $\Delta Cn$ scores for all combinations of charge, tryptic and missed cleavage state. For searches against random and Markov databases, FP identifications were estimated by multiplying the number of decoy hits by a factor of 1.6 rather than 2 to compensate for the 63% bias demonstrated in **b**. Data plotted for random and Markov databases represent the averaged results from three independent decoy database preparations under each decoy model. Error bars denote the maximum and minimum values obtained from the triplicate searches.

This work focused on estimating the frequencies of FP peptide identifications. As the basis for protein identifications, peptide FP rates necessarily influence protein FP rates. This is most clearly demonstrated by the common observation that most incorrect peptide identifications are ones in which the corresponding protein was identified by just that one peptide. Because of this, many researchers routinely discard single peptide identifications to achieve estimated FP rates approaching 0%. It must be pointed out, however, that these single-protein identifications, which after filtering can be mostly correct, often represent 30–50% of a processed data set. Their removal disproportionably reduces the sensitivity of the analysis. Instead, we advocate that decoy hits be used to design more stringent criteria for single-peptide identifications.

Ideally, each peptide would map to only one protein. Measuring the protein FP rate would then be trivial given the peptide FP rate as incorrect protein identifications would be entirely based on incorrect peptide identifications. As a result of shared sequence between protein homologs or protein isoforms, peptides can often implicate networks of related proteins, not just individual proteins. Deciphering the most likely protein sequence that gave rise to the observed peptides can be a complicated computational task (reviewed in reference 18). One limitation of the target-decoy approach is that this type of misidentification is essentially invisible and may contribute to an overestimation of the final number of correctly identified proteins. Nonetheless, from the standpoint of protein sequence networks at least, peptide FP rates can be effective in estimating protein FP rates.

The purpose of this report was to provide convincing evidence that the target-decoy approach for estimating FP identifications is an effective way to evaluate large-scale proteomics data generated by LC-MS/MS analyses. We have shown the assumptions one must make during its application are reasonable. The limitations of this approach can be projected and are manageable. This method does not replace the need for manual validation in certain smaller-scale analyses. It does, however, provide a means to harness trends in MS/MS data allowing the discovery of peptide attribute combinations that correspond with correct identifications. Using the measurements described in **Table 1**, the strength of this correspondence can be described in a formal, probabilistic fashion. When data are evaluated in this way, multiple experiments can be fairly compared and evaluated with functionally similar criteria[1]. Moreover, by adjusting criteria for specific subsets of data, such as proteins identified by single peptides, it may be possible to rescue a large portion of correct identifications that would otherwise be ignored. Because of its inherent simplicity, the essential aspects of the target-decoy strategy can be applied with minimal computational resources. It is therefore an approach that is accessible to any laboratory using any instrument platform and any database-searching algorithm to conduct mass spectrometry–based proteomics experiments.

**ACKNOWLEDGMENTS**

This work was supported in part by US National Institutes of Health (GM67945 and HG00041 to S.P.G.). We thank S. Beausoleil, P. Everley, S. Gerber and W. Haas for continuing and insightful discussions, and Sage-N for implementing our idea of the pseudo-reversed searches on their SEQUEST platform.

**COMPETING INTERESTS STATEMENT**

The authors declare that they have no competing financial interests.

**Published online at http://www.nature.com/naturemethods/**
**Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions**

1. 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* **2**, 667–675 (2005).
2. Chen, Y., Kwon, S.W., Kim, S.C. & Zhao, Y. Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J. Proteome Res.* **4**, 998–1005 (2005).
3. Eng, J.K., McCormack, A.L. & Yates, J.R., III. 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).
4. Moore, R.E., Young, M.K. & Lee, T.D. Qscore: An Algorithm for Evaluating SEQUEST Database Search Results. *J. Am. Soc. Mass Spectrom.* **13**, 378–386 (2002).
5. 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).
6. Kislinger, T. *et al.* PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol. Cell. Proteomics* **2**, 96–106 (2003).
7. Haas, W. *et al.* Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol. Cell Proteomics* **7**, 1326–1337 (2006).
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* **20**, 3551–3567 (1999).
9. Olsen, J.V., Ong, S.E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3**, 608–614 (2004).
10. Nielsen, M.L., Savitski, M.M. & Zubarev, R.A. Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry. *Mol. Cell. Proteomics* **4**, 835–845 (2005).
11. Resing, K.A. *et al.* Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **76**, 3556–3568 (2004).
12. Qian, W.J. *et al.* Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J. Proteome Res.* **4**, 53–62 (2005).
13. Higdon, R., Hogan, J.M., Van Belle, G. & Kolker, E. Randomized sequence databases for tandem mass spectrometry peptide and protein identification. *OMICS* **9**, 364–379 (2005).
14. Beausoleil, S.A. *et al.* Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. USA* **101**, 12130–12135 (2004).
15. Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P. & Gygi, S.P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **22**, 214–219 (2004).
16. 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).
17. Sadygov, R.G. & Yates, J.R., III. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **75**, 3792–3798 (2003).
18. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).
19. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292 (2006).
20. Everley, P.A. *et al.* Enhanced analysis of metastatic prostate cancer using stable isotopes and high mass accuracy instrumentation. *J. Proteome Res.* **5**, 1224–1231 (2006).
21. Kersey, P.J. *et al.* The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985–1988 (2004).