Methods

Materials and General Procedures. All chemical reagents were of analytical grade, obtained from commercial suppliers, and used without further purification unless otherwise noted. All reaction flasks were oven dried prior to use. Analytical thin layer chromatography (TLC) was performed on Analtech 250-mm silica gel G plates and visualized by staining with ceric ammonium molybdate, ninhydrin, or by absorbance of UV light at 254 nm. ^1H NMR spectra were obtained with a 400 MHz Bruker spectrometer. Chemical shifts are reported in ppm referenced to the solvent peak for ^1H NMR. Coupling constants (J) are reported in Hz. Reversed-phase (RP) HPLC was performed on a Rainin Dynamax SD-200 HPLC system with 210-nm and 254-nm detection using a Microsorb C18 analytical or preparative column.

Synthesis of Halogenated Tyrosine Salts. The halogenated tyrosine salts 5, 6, 7, and 8 were prepared according to literature procedures (1-3).

General Procedure for the Preparation of Iodoacetamide Derivatives of Halogenated Tyrosine Analogs. To the halogenated tyrosine salt (100 mg) in anhydrous DMF (0.25 mL) was added anhydrous sodium carbonate (2 equiv.) and the mixture left to stir for 0.5 h at room temperature under a nitrogen atmosphere. Chloroacetyl chloride (1 equiv.) was added dropwise over 5 minutes and the reaction was stirred at room temperature under a nitrogen atmosphere for 1 h before the reaction was transferred to a separatory funnel with ethyl acetate (15 mL). The organic layer was washed with 1 M HCl (5 mL) and the layers separated. The aqueous layer was extracted twice with ethyl acetate (5 mL) and the combined organic extracts were dried over MgSO₄ prior to removal of solvent by rotary evaporation. The crude material was then dissolved in anhydrous DMF (0.5 mL) and sodium iodide was added (6 equiv.). The mixture was stirred at rt in the dark under a nitrogen atmosphere for 24 hours before the mixture was transferred to a separatory funnel with ethyl acetate (15 mL). The organic layer was washed with water (5 mL) and the layers were separated. The aqueous layer was extracted twice with ethyl acetate (5 mL) and the combined
organic extracts were dried over MgSO₄. The crude product was purified using reverse-phase HPLC with a gradient of 15% to 60% acetonitrile containing 0.1% TFA over 40 min. All fractions were kept in foil until the solvent was removed by rotary evaporation to yield a white solid. **Compound 1.** \(^1\)H NMR (400 MHz, CD\(_3\)OD): \(\delta = 7.36\) (s, 2H), 4.59-4.55 (m, 1H), 3.73 (d, 1H, \(J = 9.6\) Hz), 3.64 (d, 1H, \(J = 9.6\) Hz), 3.11 (dd, 1H, \(J = 4.8, 14.0\) Hz), 2.89-2.83 (m, 1H).

**Compound 2.** \(^1\)H NMR (400 MHz, CD\(_3\)OD): \(\delta = 7.33\) (d, 1H, \(J = 2\) Hz), 7.03 (dd, 1H, \(J = 2, 8.4\) Hz), 6.80 (d, 1H, \(J = 8.4\) Hz), 4.57-4.54 (m, 1H), 3.72 (d, 1H, \(J = 10\) Hz), 3.65 (d, 1H, \(J = 10\) Hz), 3.09 (dd, 1H, \(J = 4.8, 14.0\) Hz), 2.88-2.83 (m, 1H).

**Compound 3.** \(^1\)H NMR (400 MHz, CD\(_3\)OD): \(\delta = 7.16\) (s, 2H), 4.56-4.53 (m, 1H), 3.73 (d, 1H, \(J = 9.6\) Hz), 3.64 (d, 1H, \(J = 9.6\) Hz), 3.11 (dd, 1H, \(J = 4.8, 14.0\) Hz), 2.89-2.83 (m, 1H).

**Compound 4.** \(^1\)H NMR (400 MHz, CD\(_3\)OD): \(\delta = 7.17\) (d, 1H, \(J = 2\) Hz), 6.99 (dd, 1H, \(J = 2, 8.4\) Hz), 6.81 (d, 1H, \(J = 8.4\) Hz), 4.58-4.54 (m, 1H), 3.72 (d, 1H, \(J = 10.0\) Hz), 3.65 (d, 1H, \(J = 10.0\) Hz), 3.09 (dd, 1H, \(J = 4.8, 14.0\) Hz), 2.89-2.83 (m, 1H).

**Barstar Mutagenesis and Purification.** A plasmid containing the *Bacillus amyloliquefaciens* protein Barstar as a 6xHis fusion in a pQE30 expression vector was obtained from D. Tirrell (California Institute of Technology). A construct encoding the single point mutant I26C was prepared using the Quickchange protocol (Stratagene).

**Forward 5’** – GGG GAA CAA ATC AGA AGT TGC AGC GAC CTC CAC CAG AC

**Reverse 5’** – GTC TGG TGG AGG TCG CTG CAA CTT CTG ATT TGT TCC CC

The mutant was expressed and purified analogously to the wild-type protein (4). The purified I26C Barstar mutant was verified by ESI-HRMS, calcd for C\(_{518}\)H\(_{802}\)N\(_{146}\)O\(_{156}\)S\(_3\) [Avg. mass]: 11667.13 Da, found 11667.2 Da.

**Barstar Labeling with Dibromide Cysteine-alkylating Tag.** A 50 µL aliquot of a 0.4 mg mL\(^{-1}\) solution of I26C Barstar in 250 mM ammonium bicarbonate was labeled with dibromide tag 1 following the same protocol for BSA labeling.

**LC-MS Analysis.** Samples were subjected to RP chromatography with an Agilent 1200 LC system that was connected in-line with an LTQ Orbitrap XL hybrid mass spectrometer. External mass calibration was performed prior to analysis. A binary solvent system consisting of buffer A (0.1% formic acid in water (v/v)) and buffer B (0.1% formic acid in acetonitrile (v/v)) was employed.

For BSA peptide samples, the mass spectrometer was outfitted with a nanospray ionization source. The LC was performed using a 100 µm fritted capillary (New Objective) pre-column self-packed with 1 cm of 5 µm, 200 Å Magic C18AQ resin (Michrom Bioresources) followed by a 100 µm fused silica capillary (Polymicro Technologies) self-packed with 10 cm of 5 µm, 100 Å Magic C18AQ resin (Michrom Bioresources). After injection of the sample and a 10 min loading step in 2% buffer B, a gradient was employed from 10% to 40% buffer B for 62 min., followed by a washing step in 99% buffer B for 10 min. A solvent split was used to maintain a flow rate of 200 nL min\(^{-1}\) at the column tip. Mass spectra were recorded in a single stage of MS in positive ion mode over the m/z scan range of 400 to 2000 using the Orbitrap mass analyzer in full-scan, profile mode, at a resolution of 60,000 (at 400 m/z).
For the intact Barstar samples, the mass spectrometer was outfitted with an Ion Max electrospray ionization source. The LC was equipped with a Poroshell 300SB-C8 column and a 100 µL sample loop. For each run, after 100 to 200 picomoles of protein was injected onto the column, analyte trapping was performed for 5 min with 0.5% B, followed by a linear gradient from 30% to 95% B over 19.5 min, and finished with a washing step in 95% B for 5 min. The flow rate was maintained at 90 µL min⁻¹. Mass spectra were recorded in positive ion mode over the m/z scan range of 500 to 2000 using the Orbitrap mass analyzer in full-scan, profile mode. Raw mass spectra were processed using Xcalibur (version 4.1, Thermo) and measured charge state distributions were deconvoluted using ProMass (version 2.5 SR-1, Novatia), using the default “small protein” parameters and a background subtraction factor of 1.5.

**Full-scan Data Analysis and Inclusion List Generation.** All full-scan LC-MS data were collected in profile mode. Noise reduction and peak detection were performed using software developed in-house based on the method described by Du et al. which makes use of a continuous wavelet transform (5). The centroided mzXML files were then searched for the presence of a desired isotopic pattern using software developed in-house following the algorithm described in the main text. The output was an inclusion list that contained the m/z values and a retention time window of three minutes for each tagged species.

**True Positive Determination.** This was done by performing an in silico digestion of BSA (up to two tryptic missed cleavages and a mass range of 600 to 4000 Da), and predicting the masses of tag labeled single cysteine-containing peptides. From this list of putative masses, LC-MS data from samples composed of 30 picomoles of halogen-tagged BSA were manually searched by obtaining an extracted ion chromatogram (EIC) for each predicted mass, in up to five charge states. Each EIC was used to determine the visual presence of the appropriate halogen pattern. Peptides were considered a true positive if at least one charge state of the predicted mass value with the appropriate isotopic signature was detected (see Supplementary Table 4).

**LC-MS/MS Analysis.** Samples were subjected to RP chromatography with an Agilent 1200 LC system that was connected in-line with an LTQ Orbitrap XL hybrid mass spectrometer. External mass calibration was performed prior to analysis. A binary solvent system consisting of buffer A (0.1% formic acid in water (v/v)) and buffer B (0.1% formic acid in acetonitrile (v/v)) was employed. The mass spectrometer was outfitted with a nanospray ionization source. The LC was performed using a 100 µm fritted capillary (New Objective) pre-column self-packed with 1 cm of 5 µm, 200 Å Magic C18AQ resin (Michrom Bioresources) followed by a 100 µm fused silica capillary (Polymicro Technologies) self-packed with 15 cm of 5 µm, 100 Å Magic C18AQ resin (Michrom Bioresources). After injection of the sample and a 20 min. loading step in 2% buffer B, a gradient was employed from 10% to 35% buffer B for 120 min, followed by a washing step in 99% buffer B for 15 min. A solvent split was used to maintain a flow rate of 200 nL min⁻¹ at the column tip.

Iodoacetamide-labeled BSA samples were subjected to a data-dependent acquisition method, where each full-scan mass spectra was recorded in positive
ion mode over the m/z scan range of 400 to 1800 using the Orbitrap mass analyzer in profile mode at a resolution of 60,000 (at 400 m/z). The ten most intense peaks were selected for collision induced dissociation (CID) fragmentation in the linear ion trap. Dynamic exclusion and charge state screening were enabled, rejecting ions with an unknown or +1 charge state. An isolation window of 3, a threshold of 500 ion counts, and activation energy of 35 were used when triggering a fragmentation event.

Dibromide-labeled BSA samples were subjected to an inclusion-list driven targeted acquisition method. The first step was collecting only full-scan mass spectra in positive ion mode over the m/z scan range of 400 to 1800 using the Orbitrap mass analyzer in profile mode at a resolution of 60,000 (at 400 m/z). The data were searched for the presence of dibromide-labeled peptides and each labeled species was recorded in an inclusion list, which contained the m/z value (M+2 ion in the labeled peptide’s isotopic envelope) and a retention time window for each labeled peptide. The same sample (and same injection volume) was then reanalyzed using an inclusion-list driven selection of precursor ions for fragmentation. The inclusion list was entered into the global mass list, and for each full-scan mass spectra up to eight CID fragmentation events could take place in the linear ion trap. Dynamic exclusion and charge state screening were enabled, rejecting ions with an unknown or +1 charge state. An isolation window of 4, a threshold of 500 ion counts, and activation energy of 35 were used when triggering a fragmentation event.

**Fragmentation Data Analysis.** Raw MS/MS data were searched using Proteome Discoverer 1.2 (Sequest, Thermo-Fisher) against the human database (UniProtKB/Swiss-Prot, November 2009) with the Bovine Serum Albumin sequence appended (Uniprot P02769), allowing for three missed cleavages, one non-tryptic termini, and the following variable modifications: methionine oxidation and either cysteine carboxamidomethylation or cysteine labeling with dibromide tag 1 (allowing for all three isotopologs). Precursor ion tolerance was set to 10 ppm and fragment tolerance was set to 0.8 Da. An FDR of 1% was used for filtering iodoacetamide-labeled peptide assignments. For dibromide-labeled peptide assignments the additional high intensity isotope peaks associated with each fragment ion that contained the dibromide tag lowered Sequest’s XCorr calculation. Therefore, an FDR of ≤ 8% and a delta of 10 ppm for the precursor mass were used to narrow the search results, and each correct assignment was manually verified.
Supplemental Figure 1: Synthesis of halogenated cysteine-alkylating tags.
**Computational Simulation of Peptides With or Without a Halogen Tag**

The natural abundances of biological elements produce a skewing effect on a peptide's isotopic distribution with increasing molecular weight. This change was computationally simulated for an average peptide ranging in weight from 500 Da to 5000 Da. At each molecular weight shown, the average elemental composition for a peptide was generated without or with a halogen tag. Using the natural abundances of the pertinent isotopes, the isotopic envelope was calculated for each molecular formula.

![Graphs](image)

**Supplementary Figure 2:** The natural abundances of biological elements produce a skewing effect on a peptide's isotopic distribution. a) The change in isotopic distribution in an average peptide ranging in weight from 500 Da to 5000 Da. The change in the isotopic distribution of the same peptide with a (b) monochloride-, (c) dichloride-, (d) monobromide-, or (e) dibromide-label.
Analysis of the Averagine Model in Predicting the Elemental Composition of an Unlabeled or a Halogen-labeled Peptide of a Known Mass

The accuracy of the averagine method was confirmed by comparing the root mean squared deviation (RMSD) between the isotopic envelope calculated for the elemental compositions of 20,000 human tryptic peptides, with the isotopic envelope calculated for an averagine predicted elemental composition based on accurate molecular weight (MW), i.e. “Null” pattern. In calculating the RMSD, m/z was assumed to be identical; relative peak heights, i.e. the isotopic pattern, were compared. Under this context, the “Null” pattern had a median deviation of less than 4% from the actual pattern. If the molecular weight (MW) is incorrect by 10%, the averagine model still produces an accurate estimation of the actual elemental composition, and consequently the actual isotopic envelope. However, as we consider peptides that contain halogenated tags, the averagine prediction of elemental composition from a given MW becomes less accurate as we proceed from a monochloride-, to a dichloride-, to a monobromide-, and finally to a dibromide-labeled peptide. The consequence of this deviation from the averagine model is visually depicted in Supplementary Figure 2, where the isotopic envelope associated with a halogen labeled peptide is markedly different from that of an unlabeled peptide.

Supplementary Figure 3: The averagine model predicts peptide elemental composition. A boxplot of the RMSD intensity difference between the isotopic envelope calculated based on the actual elemental compositions of 20,000 human tryptic peptides with their averagine predicted elemental compositions (“Null” pattern) based on MW, or with an error of ±10% in MW, or with the addition of the indicated halogen tag. The bottom and top of the red box represent the 25th and 75th percentile of data, respectively, the black bar in the red box is the median, and the ends of the whiskers represent the full range of data, except for outliers, which are depicted as open circles.
Schematic of the Graph-theoretic Construct

In many cases, the searching algorithm has to match isotopic signatures of halogen-labeled peptides in non-ideal data, which can produce a significant number of false positive identifications. To reduce this number, the algorithm exploits two features of LC-MS data: peptides are often detected in multiple charge states and in several adjacent scans. If a given match is really from a peptide, then it is likely that more of the same peptide will be found in neighboring scans and in multiple charge states. Both of these factors were taken into account using a graph-theoretic construct. A visual representation of this technique is illustrated in Supplementary Figure 4.

Supplementary Figure 4: A schematic of the graph-theoretic construct. (a) Each dot represents a putative peptide match based solely on pattern matching, many of which are false positives. (b) Edges are drawn between dots (i.e. nodes) that could represent the same peptide. (c) The graph is decomposed into disjoint sets, which can be scored and filtered to remove false positive matches. The simplest filter is to discard any nodes that have no neighbors. In practice, a multi-tiered scoring approach works well, where each set of nodes is awarded points depending on the charge states detected, the total number of nodes in the set, and the pattern alignment score of the best match in each set.
A Computational Simulation to Analyze the Effect of a Graph-theoretic Construct on the False Positive Rate

A Monte Carlo simulation was performed to analyze the ability of a graph-theoretic construct to improve discrimination between real and false positive hits. The graph geometry used is similar to the schematic in Supplementary Figure 4, where nodes are placed at regular intervals on a $C \times 100$ grid, where $C$ is the number of charge states and 100 scans are considered. Nodes were considered to be neighbors if they were separated by one scan or less, e.g. nodes in scan 1 and 3 are neighbors, while nodes in scan 1 and 4 are not.

Using the assumption that each node has an independent probability of being assigned a positive value (the individual match rate), a Monte Carlo simulation was performed by randomly assigning nodes a binary value of either true or false (“matched” or “not-matched,” respectively) with a probability $p$, except for a reference node at scan 0, which was given a charge state +1 and defined as true. Edges were then drawn between all neighboring nodes assigned a value of true, and the final graph was decomposed into disjoint sets. The set containing the reference node was then analyzed to determine the size of the set, and a net match rate was determined based on a minimum required number of nodes in the set.

The Monte Carlo simulation was performed for values of $C$ in the range $[0,6]$ and values of $p$ in the range $[0,1]$, at intervals separated by 0.001 units, with 500,000 replicates at each value of $p$. The results (Supplementary Figure 5) show that the graph-theoretic construct is able to strongly amplify differences between the true positive and false positive determination rates, leading to a higher confidence in true positives and a lower incidence of false positives.
Supplementary Figure 5: A graph-theoretic construct can strongly amplify the difference between the true positive and false positive rates of a naïve pattern-searching algorithm. In this analysis, the net match rate for a single peptide (the probability that the peptide is a true positive given that it should be found in multiple charge states and in multiple scans, i.e. using the graph-theoretic approach) was analyzed as a function of the individual match rate (the probability that the peptide is a true positive when treated as an individual, isolated isotopic pattern match, i.e. without using the graph-theoretic). Monte Carlo simulations were performed on the graph construct under searching parameters allowing up to (a) 3, (b) 4, or (c) 5 charge states of a matched peptide within two scans, in each case varying the number of neighbors required before a match is considered valid. (d) The approach was also analyzed by fixing the minimum number of neighbors at 15 and comparing the effects of allowing up to 3 to 6 identified charge states. As can be seen, when the individual match rate was low (less than ~0.3), which was true of false positives identified by the naïve pattern-searching algorithm, the match received a reduced rate when using the graph, as reflected in a low net match rate. Contrastingly, when the individual match rate was greater than ~0.3, which was true of true positives identified by the naïve pattern-searching algorithm, the match received an improved rate when using the graph, as reflected in a higher net match rate.
**The Dibromide Tag is Superior with Respect to the False Negative Rate**

To determine the false negative (FN) rate for each halogen tag, we used the same searching parameters that were used for calculating sensitivity (i.e. fixed the maximum number of false positive identifications at 100 when searching MS data derived from 10 µg of Jurkat whole cell lysate without BSA), since

\[
\text{FN} = 1 - \text{sensitivity}
\]

To calculate sensitivity:

\[
\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{TP} = \text{true positives}
\]

For this analysis, the sum of TP + FN is defined as the total manually detectable single-cysteine containing peptides in a sample of 30 pmoles of halogen-tagged BSA. Therefore, we are under approximating our sensitivity at low BSA concentrations because we are assuming that all the possible detectable peptides at the high concentration are also detectable at the low concentration. This assumption was made because we wanted to calculate the sensitivity associated with the entire process, including effects of sample dilution and MS sensitivity limitations. The calculated false negative rates are shown in Supplementary Figure 6.

**Supplementary Figure 6**: The dibromide motif is superior to other halogen motifs with respect to the false negative rate. By titrating 3.0 to 0.03 picomoles of halogen-tagged BSA into 10 µg of Jurkat whole cell lysate and subjecting the tryptic digest to LC-MS, when searching the datasets the dibromide tag resulted in a lower false negative rate at all concentrations.
The Dibromide Tag Maximizes Sensitivity and Selectivity in ROC plots

In order to generate ROC (Receiver Operating Characteristic) plots for each halogen tag, we made the following assumptions:

1. Total positives (P) were defined for each sample of halogen-labeled BSA based on the total expected peptides to be found, as defined from manual searching of 30 picomoles of BSA in the absence of lysate (see Supplementary Table 4). This was 25 for dibromide-, 19 for dichloride-, and 24 for monobromide-labeled BSA.

2. True positives (TP) were determined from a computational search for each single cysteine-containing BSA peptide from searching a sample containing the indicated quantity of halogen-tagged BSA in the presence of 10 µg of Jurkat whole cell lysate. In order to vary the TP values we adjusted tolerances for key parameters of the graph-theoretic algorithm.

3. Total Negatives (N) were defined as the total number of peptides in the digested sample that are not labeled with a halogen tag (i.e., the number of peptides in trypsinized Jurkat whole cell lysate). For this calculation we approximated that there are 35,000 proteins in the human genome and that we would expect a total of 10 tryptic peptides per protein, giving 350,000 total peptides.

4. False Positives (FP) were determined as the difference between all ions that were computationally identified and the number of true positive ions. Unfortunately, we could not convert the number of ions into the number of peptides because we don’t know what peptides each ion represents. In contrast, when determining the TP we could convert ions into peptides because we were only considering peptides from BSA. In order to vary the FP values we adjusted tolerances for key parameters of the graph-theoretic algorithm.

Therefore, our ROC plots are generated as

\[
\text{True Positive Rate (TPR)} = \frac{\text{TP}_{\text{peptides}}}{P_{\text{peptides}}},
\]

\[
\text{False positive rate (FPR)} = \frac{\text{FP}_{\text{ions}}}{N_{\text{peptides}}},
\]

which implies that the calculated FPR is an over approximation (since we report ions/peptides instead of peptides/peptides), and the actual FPR is expected to be better than reported.

As illustrated in Supplementary Figure 7, the ROC plots demonstrate that the IsoStamp algorithm can be tailored for each halogen tag, optimizing sensitivity (high TPR) and selectivity (low FPR) in each case. Additionally, for each halogen tag, the FPR is low while the TPR is high, with the dibromide tag showing the best overall performance, even at low concentrations.
Supplementary Figure 7: Receiver operating characteristic curves (ROC) for the IsoStamp algorithm. Each plot shows the results of varying key parameters of the graph-theoretic algorithm (to adjust sensitivity and specificity) and then searching samples containing the indicated quantity of either (a) dibromide-, (b) dichloride-, or (c) monobromide-labeled BSA peptides in the presence of 10 µg of Jurkat whole cell lysate.
Analysis of Barstar Labeling with the Averagine Model

To determine whether the dibromide tag’s isotopic signature can be detected on large peptides, we labeled the small protein Barstar from B. cenocepacia (11.7 kDa, including a N-terminal 6xHis tag) with dibromide tag 1 on a single cysteine residue introduced by site-directed mutagenesis (I26C). The labeled intact protein was analyzed by LC-MS on an LTQ-Orbitrap XL mass spectrometer. Shown in Supplementary Figure 7 are the mass spectra of unlabeled (panel A) and dibromide-labeled (panel B) Barstar in the +9 charge state. Using the averagine system, we predicted the mass envelope of the protein with and without the dibromide tag and depicted the peak intensities in the form of red and blue curves, respectively. An overlay of the two curves suggested that addition of the dibromide tag should cause a detectable widening of the mass envelope. In order to score how well the averagine predictions matched the experimental data, we calculated the RMSD between the peak intensities from the MS data and each of the predicted mass envelopes.

The averagine-based models for isotopic envelopes of Barstar labeled with or without the dibromide tag 1 were predicted based on the molecular weight of Barstar (monoisotopic mass of 11659.84 Da). Models (reference pattern, R) were fit against centroided experimental data by first determining the value of a scaling factor k such that the reference pattern could be optimally aligned with the data (D) in the intensity dimension. This was accomplished by performing a binary search such that the sum of the squared difference (SSD) between each peak in the reference pattern ($r_i \in R$) and its counterpart in the actual data set ($d_i \in D$) is minimized:

$$SSD = \sum_i (d_i - kr_i)^2$$

After intensity alignment, the score for the fit was calculated as:

$$Score = \prod_{i=1}^{N} f\left(\frac{|d_i - kr_i|}{\sigma \sqrt{2}}\right)$$

where $\sigma$ is a measure of peak intensity variance and $f$ is a scoring function for each peak that produces a value in the range [0,1]

$$f(x) = (1 - \varepsilon)\text{erfc}(x) + \varepsilon, \quad 0 < \varepsilon << 1$$

in which $\text{erfc}(x)$ is the complement of the Gaussian error function and the parameter is a measure of the tightness of the peak matching in the intensity dimension. In the case of measuring goodness of fit for the Barstar data, the score was calculated using values of $\varepsilon = 10^{-5}$, $\sigma = 25\%$ (RMS intensity), and $N = 16$. Because scores produced in this manner are typically very small, the log values of the scores were compared:
The experimental data for the dibromide tagged protein showed a lower RMSD from its averagine-predicted mass spectrum than from the predicted spectrum of the untagged protein (and vice versa). The small RMSD difference, while computationally detectable, suggests that masses approaching 12 kDa are most likely at the limit of detection for the IsoStamp technology. Alternative isotopic perturbations as well as a more sophisticated searching algorithm can be explored towards developing a more robust system that is capable of identifying changes in this mass region.

**Supplementary Figure 8.** The dibromide tag can be detected on purified small proteins. The averagine predicted isotopic envelope with (red) and without (blue) the dibromide tag fit to mass spectra of (a) barstar and (b) dibromide-labeled barstar. Barstar’s predicted MW = 11.7 kDa. Spectra shown are from the +9 charge state.
Detection of Dibromide-labeled BSA peptides on a Waters QTof Instrument

The ability to detect the dibromide pattern depends primarily on two instrument parameters: the $m/z$ resolution and the desired sensitivity. It is possible to match the dibromide pattern on lower resolution instruments provided that the isotopic pattern itself is recreated faithfully. Alternatively, pattern matching on a higher mass-resolution instrument requires less fidelity in the isotopic pattern itself. The type of instrument that can be used is also dependent on the acceptable false positive rate, as higher acceptable false positive identifications will permit the use of lower resolution instruments.

To test the sensitivity of detecting dibromide-labeled BSA at resolutions lower than 60,000, we performed a titration experiment with dibromide-labeled BSA (identical to that described in the Methods section), and subjected the peptide mixture to LC-MS analysis on a Waters QTof mass spectrometer a resolution of 10,000. The data were searched for dibromide-labeled peptides and the results are shown in Supplementary Figure 8.

Supplementary Figure 9: The dibromide motif can be detected on different types of mass spectrometers. A comparison of the sensitivity of detecting dibromide-labeled BSA peptides was performed between a Waters QTof (square) and a LTQ Orbitrap XL (circle) mass spectrometer. Sensitivity was determined by titrating 3.0 to 0.03 picomoles of dibromide-tagged BSA into 10 µg of Jurkat whole cell lysate and analyzing the digest by LC-MS. The number of peptides corresponding to dibromide-tagged BSA is reported as a proportion of the detectable cysteine containing peptides, and at all concentrations both instruments perform almost identically.
References
1. Ding, W., Zhang, J., Yao, Z., Lu, R., Wu, D., Li, G., Shen, Z., Sun, Y., Lin, G., Wang, C., Zhao, M., and Peng, S. (2004) The synthesis, distribution, and anti-hepatic cancer activity of YSL, Bioorg Med Chem 12, 4989-4994.
2. Prieto, M., Mayor, S., Rodriguez, K., Lloyd-Williams, P., and Giralt, E. (2007) Racemization in suzuki couplings: a quantitative study using 4-hydroxyphenylglycine and tyrosine derivatives as probe molecules, J Org Chem 72, 1047-1050.
3. McCubbin, J. A., Maddess, M. L., and Lautens, M. (2006) Total synthesis of cryptophycin analogues via a scaffold approach, Org Lett 8, 2993-2996.
4. Beatty, K. E., Xie, F., Wang, Q., and Tirrell, D. A. (2005) Selective dye-labeling of newly synthesized proteins in bacterial cells, J Am Chem Soc 127, 14150-14151.
5. Du, P., Kibbe, W. A., and Lin, S. M. (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics 22, 2059-2065.
Supplementary Table 1: The natural abundances of stable isotopes of elements found in a standard amino acid primarily exist as light isotopes, while bromine and chlorine have high abundances of their heavy isotopes.

| Element | Stable Isotopes | Natural Abundances |
|---------|----------------|--------------------|
| Hydrogen | $^1$H | 99.99% |
|         | $^2$H | 0.01% |
| Carbon  | $^{12}$C | 98.93% |
|         | $^{13}$C | 01.07% |
| Nitrogen | $^{14}$N | 99.64% |
|         | $^{15}$N | 00.36% |
| Oxygen  | $^{16}$O | 99.76% |
|         | $^{17}$O | 00.04% |
|         | $^{18}$O | 00.20% |
| Phosphorus | $^{31}$P | 99.99% |
| Sulfur  | $^{32}$S | 94.99% |
|         | $^{34}$S | 04.25% |
| Bromine | $^{79}$Br | 50.69% |
|         | $^{81}$Br | 49.31% |
| Chlorine | $^{35}$C | 75.76% |
|         | $^{37}$C | 24.24% |
Supplementary Table 2: The “averagine” peptide. Numbers shown are the number of atoms of each element type per Da of peptide. These numbers allow the elemental composition of a peptide to be predicted from its molecular weight.

| Element   | Number per Da peptide (± std. dev.) |
|-----------|-------------------------------------|
| Hydrogen  | 0.07100 ± 0.00470                   |
| Carbon    | 0.04350 ± 0.00520                   |
| Nitrogen  | 0.01260 ± 0.00180                   |
| Oxygen    | 0.01380 ± 0.00170                   |
| Sulfur\(^b\) | 0.00037 ± 0.00052                   |

\(^a\)Calculated on a random selection of 5000 human tryptic peptides with a sequence length greater than 6 amino acids.

\(^b\)Because of its rarity in peptides, there is a high error associated with predicting the number of sulfur atoms in relatively small peptides.
Supplementary Table 3: The computational detection rate for each single cysteine-containing BSA peptide from searching a sample containing the indicated quantity of halogen-tagged BSA in the presence of 10 µg of Jurkat whole cell lysate. Numbers in each column indicate the number of ions computationally found for each peptide listed (where the ion could be in the +1 to +5 charge state), while values of “NA” indicate that ions corresponding to that peptide were not expected to be found because they were not identified from manual searching of 30 picomoles of BSA in the absence of lysate (see Supplementary Table 4). Ions found were compared against a pure BSA sample subjected to tryptic digestion and analyzed by LC-MS/MS as a control for retention time and ionizability. Data were obtained on an LTQ Orbitrap XL (resolution = 60,000) or Waters QToF Premier (resolution = 10,000) mass spectrometer.

Note: Only single cysteine-containing peptides were evaluated because if a peptide contains more than one cysteine than it can contain more than one tag. This complicates our optimized searching parameters for each halogen tag because, for example, a peptide labeled with two monochloride tags will be almost identical to the same peptide labeled with only one dichloride tag. Therefore, to eliminate these complications from our analysis, we only searched for single cysteine-containing peptides.

See excel document.
**Supplementary Table 4**: The manual detection rate for each single cysteine-containing BSA peptide from searching a sample containing 30 picomoles of halogen-tagged BSA. Numbers in each column indicate the number of ions manually found for each peptide listed (where the ion could be in the +1 to +5 charge state). Data were obtained on an LTQ Orbitrap XL (resolution = 60,000) or Waters QTof Premier (resolution = 10,000) mass spectrometer.

See excel document.
**Supplementary Table 5:** A comparison between the data-dependent detection rate and the directed detection rate for each single cysteine-containing BSA peptide from searching a sample containing the indicated quantity of tagged BSA in the presence of 10 µg of Jurkat whole cell lysate. BSA was either labeled with iodoacetamide for data-dependent analysis or dibromide tag 1 for directed analysis. Numbers in each column indicate the number of ions found for each peptide listed (where the ion could be in the +1 to +5 charge state). These ions were correlated to a peptide sequence from a database search using the Sequest algorithm and manual searching of the resulting peptide sequences to select the best matches; values of 0 indicate that the peptide was not found in the dataset after filtering and manual searching. Data were obtained on an LTQ Orbitrap XL (resolution = 60,000).

See excel document.