HI-Bone: A Scoring System for Identifying Phenylisothiocyanate-Derivatized Peptides Based on Precursor Mass and High Intensity Fragment Ions

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 2013

Impact Factor: 5.64 · DOI: 10.1021/ac303239g · Source: PubMed

analytical chemistry

Technical Note

pubs.acs.org/ac

HI-Bone: A Scoring System for Identifying Phenylisothiocyanate-Derivatized Peptides Based on Precursor Mass and High Intensity Fragment Ions

Yasset Perez-Riverol,[†,‡,#] Aniel Sánchez,[†,⊥,#] Jesus Noda,[†] Diogo Borges,[∥] Paulo Costa Carvalho,[§] Rui Wang,[‡] Juan Antonio Vizcaíno,[‡] Lázaro Betancourt,[†] Yassel Ramos,[†] Gabriel Duarte,[⊥] Fabio C.S. Nogueira,[⊥] Luis J. González,[†] Gabriel Padrón,[†] David L. Tabb,[@] Henning Hermjakob,[‡] Gilberto B. Domont,*[,⊥] and Vladimir Besada*[,†]

[†]Department of Proteomics, Center for Genetic Engineering and Biotechnology, Ave 31 e/158 y 190, Cubanacán, Playa, Ciudad de la Habana, Cuba

[‡]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, U.K.

[§]Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz-Paraná, Brazil

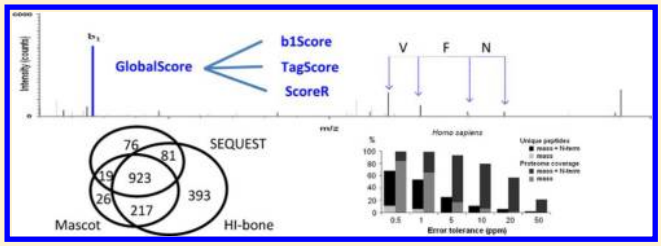
[∥]Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[⊥]Proteomics Unit, Institute of Chemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[@]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, United States

Supporting Information

ABSTRACT: Peptide sequence matching algorithms used for peptide identification by tandem mass spectrometry (MS/MS) enumerate theoretical peptides from the database, predict their fragment ions, and match them to the experimental MS/MS spectra. Here, we present an approach for scoring MS/MS identifications based on the high mass accuracy matching of precursor ions, the identification of a high intensity b1 fragment ion, and partial sequence tags from phenylthiocarbamoyl-derivatized peptides. This derivatization process boosts the b1 fragment ion signal, which turns it into a powerful feature for peptide identification. We demonstrate the effectiveness of our scoring system by implementing it on a computational tool called "HI-bone" and by identifying mass spectra of an *Escherichia coli* sample acquired on an Orbitrap Velos instrument using Higher-energy C-trap dissociation. Following this strategy, we identified 1614 peptide spectrum matches with a peptide false discovery rate (FDR) below 1%. These results were significantly higher than those from Mascot and SEQUEST using a similar FDR.



Protein identification in large-scale shotgun proteomics experiments is usually accomplished by automatically comparing theoretical mass spectra from peptides generated from a protein sequence database to those experimentally obtained typically by liquid chromatography coupled online with tandem mass spectrometry (LC−MS/MS). Examples of software tools for automatically performing this peptide spectrum matching (PSM) task are search engines such as SEQUEST,[1] Mascot,[2] X!Tandem,[3] and OMSSA.[4]

In general terms, the specificity of a PSM algorithm is inversely proportional to the peptide search space size. As such, these strategies are usually more efficient in experiments addressing model organisms that have a small and well-annotated protein sequence database derived from its genome (e.g., *Escherichia coli*). On the other hand, the current PSM algorithms can frequently use only a small number of all the generated high-quality MS/MS spectra in the experiment. The number of peptides generated after the proteolysis of complex samples still overwhelms the capacity of analysis of the most advanced LC−MS systems. As a result, unfortunately only a relatively small proportion of the acquired MS/MS spectra yields positive identifications, due either to poor spectrum quality or to insufficiently optimized scoring methods. Taken together, such aspects might significantly limit the PSM working models. These limitations motivated us to rethink how the experimental design of traditional PSM approaches is accomplished.

Here, we propose a methodology to ultimately provide increased sensitivity when analyzing phenylthiocarbamoyl-derivatized peptides (first step of the Edman degradation reaction). This derivatization process boosts the b1 fragment ion intensity and simplifies the number of fragments in the MS/MS spectrum, turning it into a powerful feature that can be

Received: November 12, 2012
Accepted: February 28, 2013
Published: February 28, 2013

   dx.doi.org/10.1021/ac303239g | *Anal. Chem.* 2013, 85, 3515−3520

used to facilitate peptide identification. As shown by Gaskell et al.,[5,6] the equivalent of a single Edman degradation stage can be performed for ions in the gas phase. The peptides are coupled in solution with phenylisotiocyanate (PITC) and ionized in either electrospray ionization (ESI) or matrix-assisted laser desorption ionization (MALDI) ion sources. After activation in the collision cell, the phenylthiocarbamoyl (PTC)-derivatized peptide dissociates specifically to yield an intense b1 fragment, consequently unlocking the possibility to determine the mass and the N-terminal residue of a given peptide in a single mass spectrum.[7,8] Wohlhueter and colleagues demonstrated that other isothiocyanate analogues bearing basic moieties can derivatize peptides and significantly improve the MS sensitivity of tagged analytes while promoting Edman fragmentation.[9] Another example of its usefulness was proposed by Yao and co-workers: they used this active chemical modification for improving the fragment ion detection in MRM (Multiple Reaction Monitoring) studies.[10]

Here, we describe a scoring system used by the tool denoted "HI-bone" and demonstrate its efficiency in an *E. coli* sample acquired on an Orbitrap Velos instrument using higher-energy C-trap dissociation (HCD). HI-bone generated scores are based on the intensity of the b1 fragment ion, and in the presence of four fragmentation patterns and partial sequence tags of the spectrum. We were able to converge to a list of 1614 PSMs and 526 proteins, using the proposed algorithm. This result is significantly better when compared to those obtained from Mascot (1185 PSMs) and SEQUEST (1099 PSMs), corresponding to 418 and 421 proteins, respectively. The HI-bone tool also reports a set of complementary subscores that can be used in the future for PSM quality assignment.

MATERIALS AND METHODS

Experimental Data and Database Search. Proteins from the *E. coli* strain W3110 cells were converted in a PTC-derivatized peptide mixture and acquired on an Orbitrap Velos instrument using HCD (see the Supporting Information). With the use of the HI-bone algorithm, the spectra file was deconvoluted and deisotoped using the "MS-Deconv" application under default parameters.[11] The HI-bone algorithm parameters used were 10 ppm tolerance at the precursor level and 5 ppm at the MS/MS level. For partial sequence tags generation, a minimum of 2 and maximum of 3 amino acids and a mass error of 0.04 m/z units was used. The resulting tryptic proteome was filtered with the number of arginine greater than 0 ($R > 0$), using the HI-bone sequence filter.

The Mascot (version 2.3.02) and SEQUEST (version 1.3) search engines and the *E. coli* subset from UniProtKB/Swiss-Prot (release 11/2011) were used to compare the HI-bone scoring system. Mascot and SEQUEST tolerance parameters were set to 10 ppm and 0.6 Da for precursor mass and product ion, respectively. Other parameters used were trypsin digestion with up to two missed cleavages, a fixed modification of 71.037 Da at cysteine residues, another fixed modification of 135.014 Da at the protein N-terminal end and lysine (PTC derivatized), and a maximum charge of +3. A shuffled decoy database was generated and joined to the target database for false discovery calculations.

A non PTC-derivatized *E. coli* data set was analyzed using HI-bone with the same configuration using above, including the modified peptides. The PTC data set was analyzed to define the score threshold for performing confident peptide identifications.

Software Description. The software used in this study implementing the HI-bone algorithm is a standalone Java application that ports a Graphical User Interface (GUI) with three main windows (The source code of the HI-bone is provided upon mail request to the corresponding author.): (i) search engine configuration parameters; (ii) peptide identification and spectrum query results; and (iii) a mass spectrum visualization panel. The search engine configuration parameters panel allows the definition of post-translational modifications, enzymes, and the MS and MS/MS error tolerances. The library ms-core-api (http://code.google.com/p/pride-toolsuite/) was used to handle different spectrum file formats and the pride-mod library (http://www.ebi.ac.uk/~maven/m2repo_snapshots/uk/ac/ebi/pride-mod/pride-mod/) was used for the handling of post-translational modifications. The spectrum visualization panel allows the manual inspection of the nonidentified and identified spectra. Using the pride-mzgraph-browser library (http://code.google.com/p/pride-toolsuite/wiki/PRIDEmzGraphBrowser), the spectrum panel can represent the spectrum fragmentation and the intense b1 peak.

RESULTS AND DISCUSSION

Fragment Scoring Functions. We have developed a novel empirical scoring scheme based on the presence of high b1 ion intensities for PTC-derivatized peptides. Following the b1 fragmentation pattern of the modified peptides, the b1 score (b1Score) represents the rank (order) of the signal assignment to the b1 ion in a sorted list by signal intensity. Rather than work with intensity directly, the software evaluates peaks by their intensity ranks. It reflects Bern's observation that the significance of fragment ions may be judged more accurately by their intensity ranks than by their relative intensities.[12] Also, it shows the relation between the different possible b1 signals assigned by fragment masses in the low mass region (193−400 m/z units):

$$\text{b1Score} = \frac{1}{O_i} \tag{1}$$

where O_i is the rank of the b1 signal assigned in a list of low mass signals sorted by intensity in a nonincreasing order. The b1Score values are in the range of $[0-1]$ and they represent the order, depending of its intensity, in which the b1 fragment was assigned to a given theoretical fragment. We observed that for PTC-derivatized peptides, based on peak presence, internal sequences starting by proline and glycine were more common than those containing other amino acids. In addition, for the cleavage at the C-terminal amino acid of an internal fragment, there was a significant preference for aspartic and glutamic acid. Similar results were generalized recently by Mann and co-workers for HCD fragmentation.[13] Then, for each peak presence of proline, glycine, aspartic, and glutamic acid, we computed the corresponding intensity score:

$$\text{Score}_R = \frac{1}{O_i} \tag{2}$$

where O_i is the rank of the signal assigned to the amino acid (proline, glycine, aspartic, and glutamic acid).

Sequence Tags Scores. The last score is derived from inferring partial sequence tags directly from observed fragment ions. Sequence tagging comprises a middle path between database searches and *de novo* strategies. Our software derives

dx.doi.org/10.1021/ac303239g | Anal. Chem. 2013, 85, 3515−3520

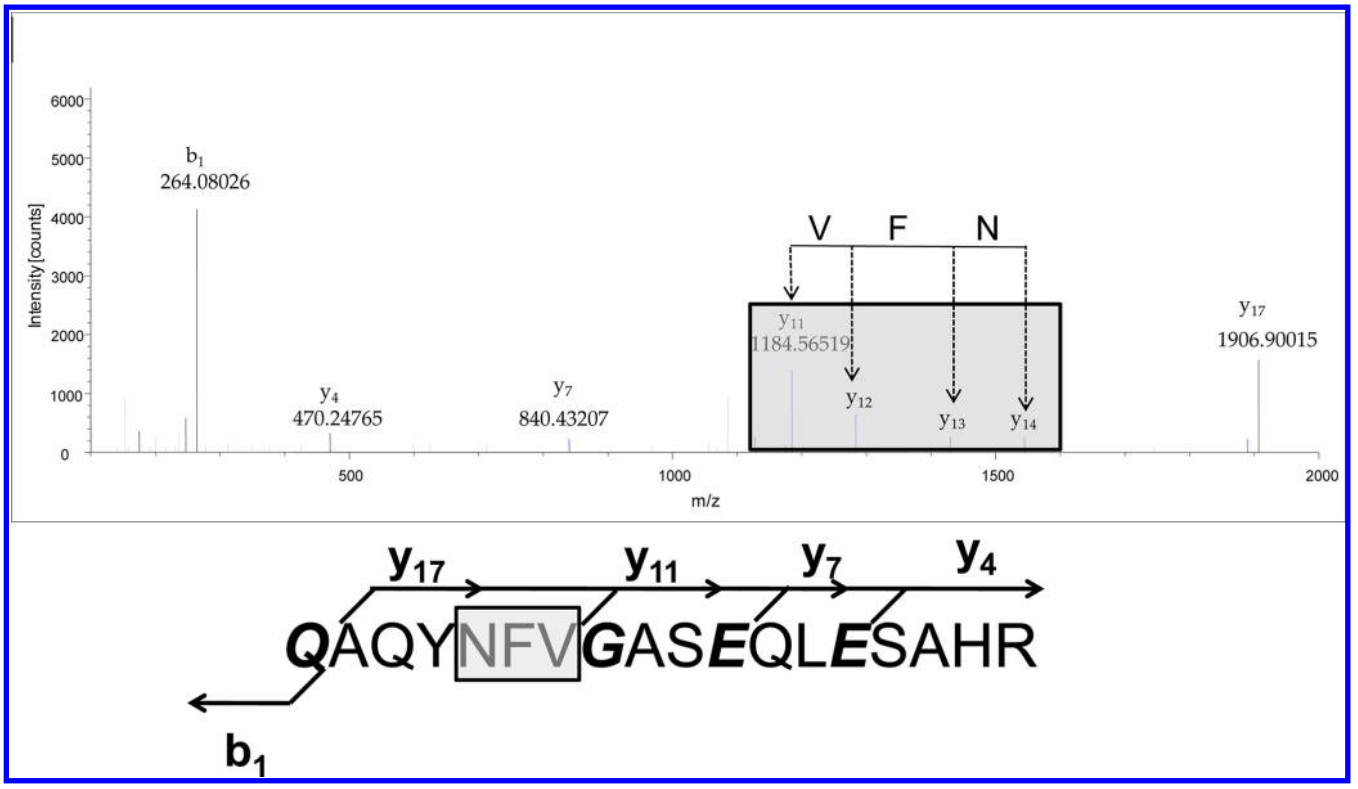


Figure 1. Example of fragmentation pattern of PTC-derivatized peptides.

partial sequence tags from a tandem mass spectrum, which are then evaluated against a protein sequence database to interpret the remainder part of the spectrum.[14−16] The HI-bone sequence tag algorithm seeks pairs of peaks that are separated by known amino acid masses. The spectrum can then be evaluated as a graph, with peaks represented by nodes and amino acid gaps between peaks represented by edges. When a set of peaks is joined by consecutive edges in this graph, the set constitutes a tag. For each sequence tag the algorithm computes the intensity score based on the rank order of the amino acid peaks:

$$\text{TagScore} = \sum_{A}^{i=0} \frac{1}{O_i} \tag{3}$$

where O_i is the rank of the amino acid peaks in the spectrum. This intensity score was also employed previously by Tabb and co-workers, together with a delta mass score.[15]

Global Score. We also define a global score made by the combination of the different scores (b1Score, residue Score_R, and TagScore). We generate a polynomial function:

$$\text{GlobalScore} = [(\sum_{R}^{i=0} \text{Score}_R) \times R^3] + \text{TagScore} \tag{4}$$

where R is the number of amino acids found in the peptide sequence candidate and also found in the following set: {proline, glycine, aspartic, and glutamic acid}. The global score boosts the contribution of the analyzed residues in combination with the partial sequence tags intensity scores. Also, it normalizes the contribution of each residue compared with partial sequence tags.

The HI-bone algorithm and Score Reporting. HI-bone stores the inferred PSM sequences into a tab-delimited file with all the previously described subscores. We also report the m/z fidelity score, which is the sum of squared errors (SSE) of the first peak m/z estimates for each tag,[15] the precursor delta mass, and the b1 fragment ion delta mass. These variables can be used in future developments to discriminate correct from incorrect PSMs. For example, the mass spectrum corresponding to the peptide QAQYNFVGASEQLESAHR (Figure 1) shows clearly that the b1 ion is the most intense fragment ion in the low mass region [$(\text{PITC-Q})_{exp}$ = 264.0806 Da and experimental mass = 264.0803 Da]. In this case, the b1Score = 1, glycine Score = 0.33, glutamic acid Score = 0.14, and TagScore = 4.83, allowing the peptide identification with a global score of 17.67.

The HI-bone algorithm starts by searching a peptide in the database within a given mass accuracy tolerance. The MS/MS signals are then sorted by intensity and the b1Score and the signal scores for glycine, proline, glutamic, and aspartic acid. For those peptides with a b1Score above confidence, the partial sequence tags are generated. Finally, the global score is computed for each PSM and all the scores are listed in the tab output (Figure 2).

Evaluation of the Scoring Systems. *Threshold Definition of the b1Score.* A sample from *E. coli* containing only non-PTC-modified peptides was used as a negative control, to select a confident b1Score threshold. The LC−MS/MS run and the HI-bone identification workflow were performed using the same protocol previously used for the PTC-derivatized peptide analysis. Only 0.04% of the PSMs were identified with a b1Score $\geq$ 0.2. In other words, using this confident score threshold (b1Score is >0.2), we were able to avoid false positive assignments associated with non-PITC fragments ions.

Sub-Scores Evaluation and Identification Results. Figure 3 shows a theoretical analysis of unique peptide and protein coverage, using only the precursor mass, the N-terminal residue, and different mass accuracies for *Homo sapiens* and *E. coli*. Similar to a previous study,[8] the number of unique peptides

dx.doi.org/10.1021/ac303239g | *Anal. Chem.* 2013, 85, 3515−3520

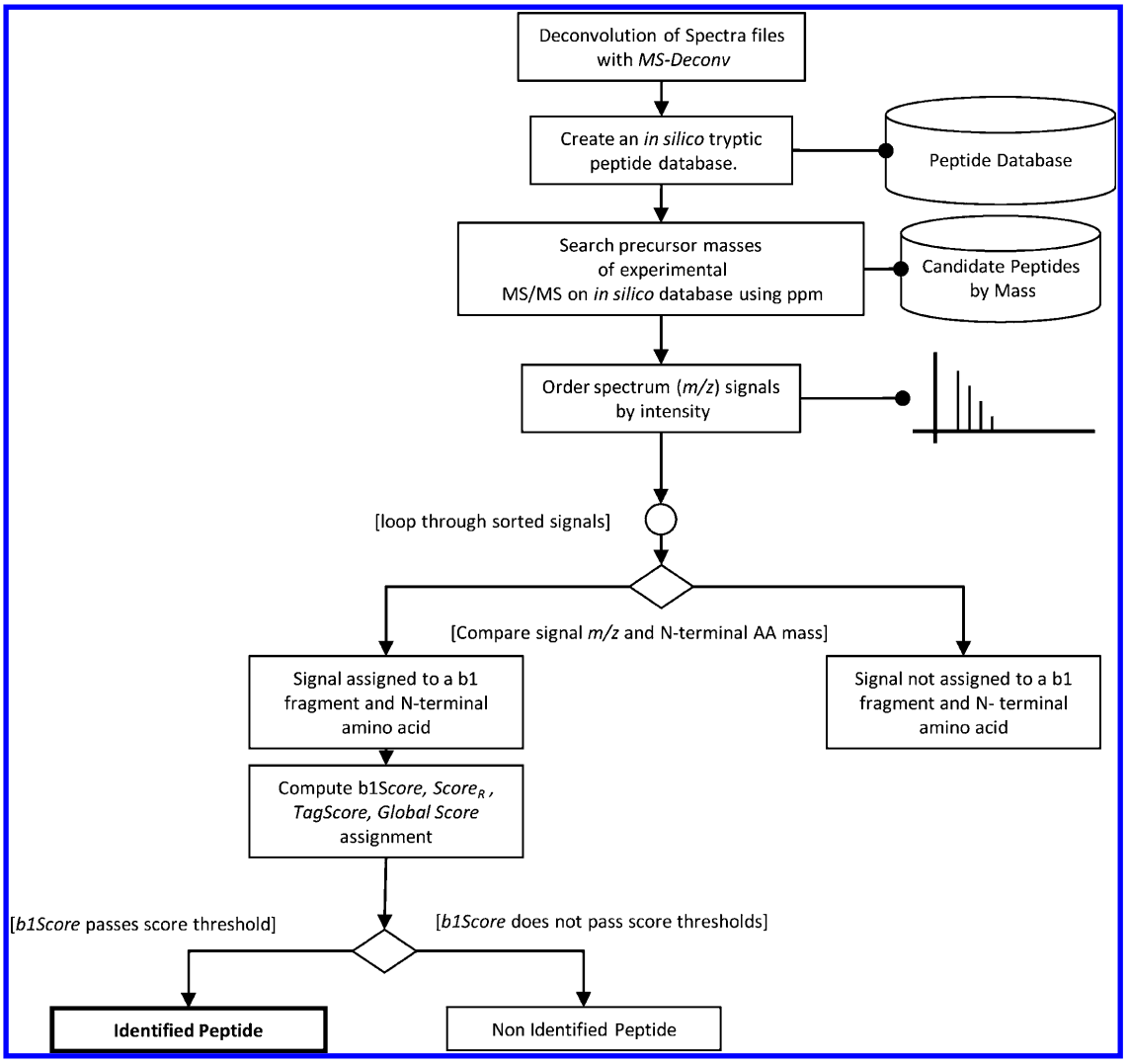


Figure 2. HI-bone algorithm workflow.

for an *E. coli* proteome is 40% for low parts per million accuracy values (1 ppm). The results for 10, 20, and 50 ppm are very poor when the precursor mass is only considered. The number of unique peptides increases when the N- terminal residue is identified for both proteomes [(A) *E. coli* and (B) *H. sapiens*] and all mass accuracies. The human proteome coverage (Figure 3C) increases considerably when the precursor mass and N-terminal residue is employed in the poor accuracy region (10, 20, and 50 ppm). This theoretical analysis opens the possibility to a new methodology of peptide and protein identification, using the precursor mass and b1 fragment ion even for low mass accuracy instruments.

The generated subscores ($Score_R$, TagScore) and the GlobalScore were explored independently using the *E. coli* data set (Figure 4A). Receiver operating characteristic (ROC) curves can be used to determine the value of scoring functions.[17] The vertical position of the point gives the number of true peptide spectrum matches passing the FDR threshold, while the horizontal position indicates the number of false peptide spectrum matches over the threshold. An ideal algorithm scores all true PSMs higher than all false ones. Therefore, a ROC plot for such an algorithm would be a right angle. Even when each subscore allows the identification of peptides by itself, the combination in a GlobalScore guarantees a higher number of identifications at a low peptide FDR.

Figure 4A shows a 1% FDR line that represents the sensitivity in PSM assignments among the various scoring functions. At 1% FDR and in consideration of the b1Score $\geq$ 0.2, the number of PSMs increase from 1485 (without b1Score threshold) to 1630 PSM. When the b1Score threshold is applied, the HI-bone score clearly outperforms the results of GlobalScore and the other subscores. GlobalScore shows the best results compared with TagScore (1372 PSMs) and $Score_R$ (1461 PSMs) independently. The use of partial sequence tags of 2 and 3 residues long worked best for increasing the TagScore sensitivity. Similar to previously obtained results,[15] the use of an intensity rank subscore (TagScore) was enough to score partial sequence tags.

Comparison with PSM Algorithms. Our results indicated that arginine-containing peptides were 99% of our confident PSMs. These findings are aligned with the fact that PITC derivatization increases arginine containing peptides, which works to our advantage by simplifying a complex peptide mixture by about 50%. Nevertheless, these arginine-containing peptides are still representative of the total protein content.[18] The simplification of a peptide mixture allows the reduction of

dx.doi.org/10.1021/ac303239g | *Anal. Chem.* 2013, 85, 3515−3520

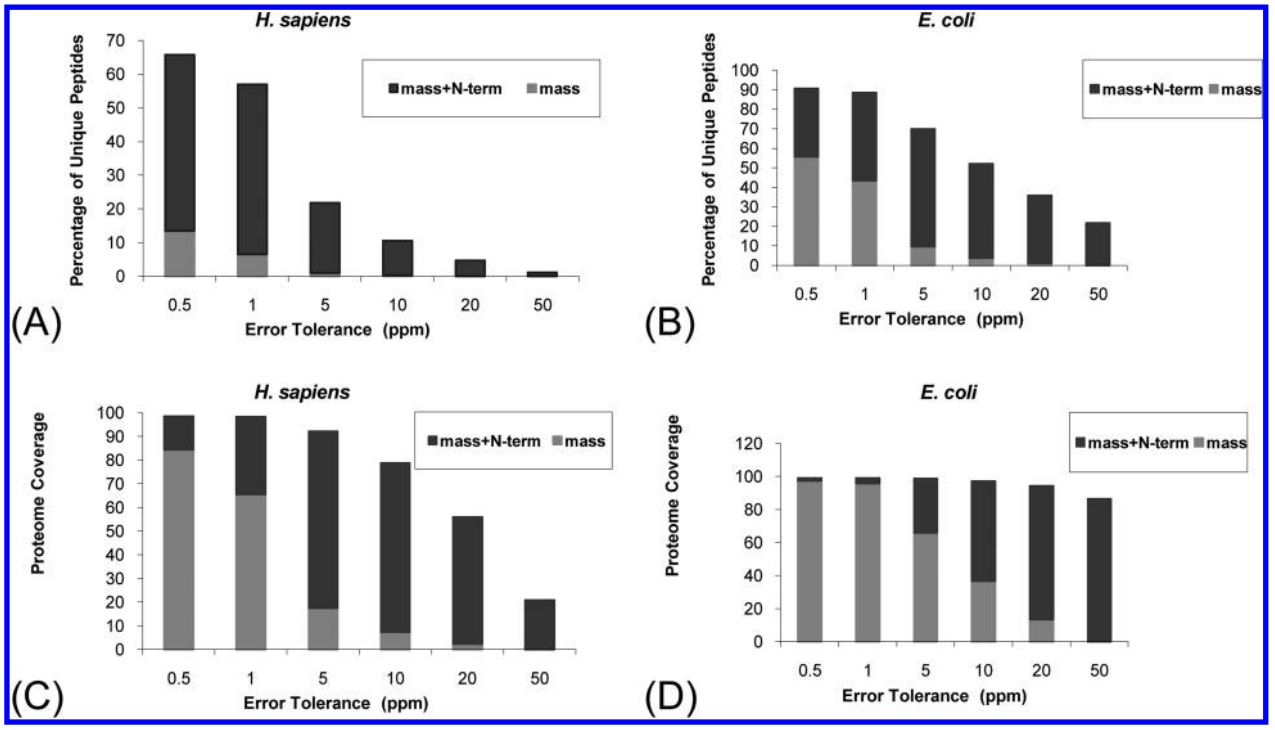


Figure 3. Theoretical identification of (A and B) unique peptide and (C and D) protein coverage, using only the precursor mass and the N-terminal residue for different mass accuracy. (A and B): Percentage of unique peptides *H. sapiens* and *E. coli* proteome, respectively. (C and D): Percentage of proteome coverage for *H. sapiens* and *E. coli*, respectively.

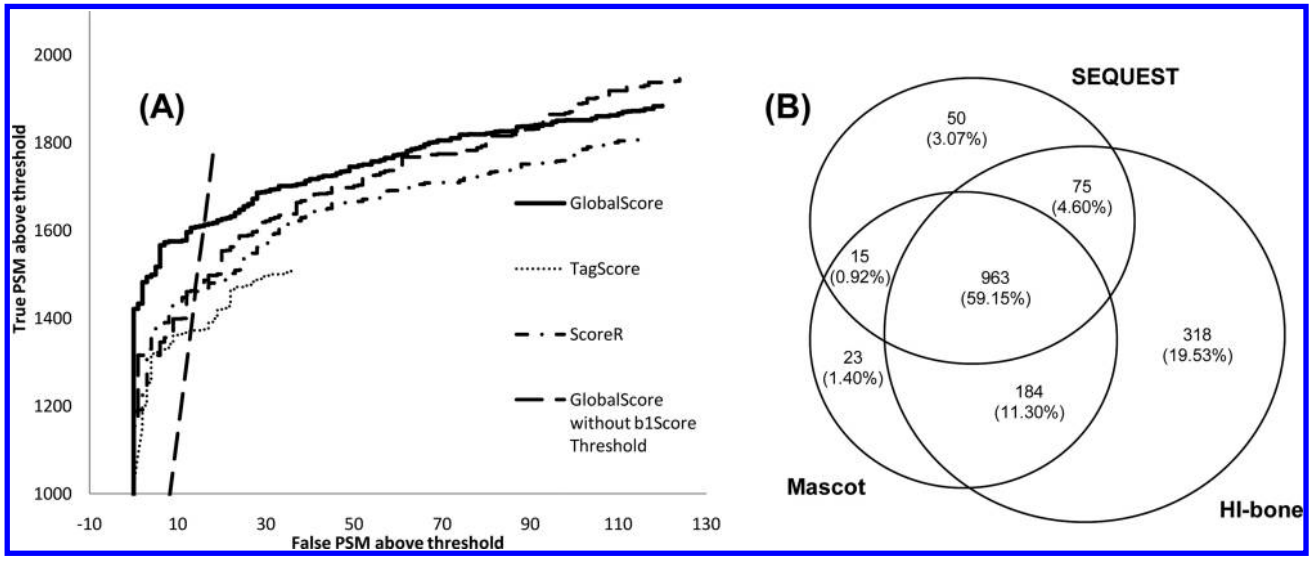


Figure 4. (A) ROC curve comparing the number of true positives to the number of false positives for various peptide false discovery rate cutoffs between 0 and 6%, for a particular scoring function (GlobalScore, TagScore, $Score_R$, GlobalScore, without b1Score threshold). (B) Venn diagram showing the distribution of PSMs, using HI-bone, Mascot, and SEQUEST.

the "database space" for HI-bone searching and increases the probability of finding unique peptides.

The number of PSMs (Figure 4B), and consequently proteins (Table 1), identified using HI-bone for PTC-derivatized PSMs, were higher than those obtained with other conventional search engines.

Figure 4B presents a Venn diagram of the PSM distribution obtained by Mascot, SEQUEST, and HI-bone. In particular, the number of PSMs identified by HI-bone was 1614 (and 526 proteins). Approximately, 22.65% (318 proteins) of the PSMs were identified exclusively with HI-bone, whereas 53% were shared between the three search engines. It is important to note that the HI-bone software was designed exclusively for the identification of phenylisothiocyanate-derivatized peptides based on the mass accuracy of the most common fragments

Table 1. Number of Proteins Identifications from PITC Peptides using HI-bone, Mascot, and SEQUEST, considering a Peptide FDR < 1%

	Mascot	SEQUEST	HI-bone
no. identified protein	418	421	526

dx.doi.org/10.1021/ac303239g | *Anal. Chem.* 2013, 85, 3515−3520

occurring for these peptides, instead of Mascot and others that have very good results with nonmodified samples. Mass spectra of PTC-derivatized peptides shift intensity to the b1 fragment ion, and this may lead to disadvantages for Mascot and SEQUEST as they were tailored to consider different intensity distributions. The HI-bone application and scoring system are devoted to PTC-derivatized or other isothiociane-derivate peptides, where the b1 fragment ion is favored and appears as the most intense fragment in the low mass region of the spectrum.[9,19] Also, the HI-bone algorithm could be applicable to the analysis of samples acquired in other mass spectrometers, such as QTOF and/or FT-ICR, ideally after internal mass calibration to increase the mass accuracy.

CONCLUSIONS

The HI-bone software outperformed Mascot and Sequest in the analysis of PTC-derivatized peptides in an *E. coli* sample. The GlobalScore scoring system takes advantage of the presence of the most frequent fragments in the mass spectrum and partial sequence tags and discards false positives generated by random b1 fragments in peptides.

ASSOCIATED CONTENT

Supporting Information

Experimental Data Description. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*V. B.: e-mail, vladimir.besada@cigb.edu.cu; G.B.D.: e-mail, gilberto@iq.ufrj.br; Tel, 53-7-2718008; Fax, 53-7-2736008.

Author Contributions

#These authors have contributed equally to the work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank the INSPUR Company from China for the kind donation of the computer cluster TS10000, used for all the calculations, and the tools developed in this manuscript. R.W. is supported by the BBSRC "PRIDE Converter" grant [reference BB/I024204/1]. J.A.V. is supported by the EU FP7 grants, LipidomicNet (Grant 202272), and ProteomeXchange (Grant 260558). G.B.D. thanks CNPq. P.C.C. thanks Fiocruz−PDTIS, CNPq universal, and CDTS for financial support. The authors want to thank Tao Xu and Professor John Yates from Scripts Laboratories for their kind support of GutenTag.

ABBREVIATIONS

ESI: electrospray ionization
FDR: False Discovery Rate
GUI: Graphical User Interface
HCD: Higher-energy C-trap dissociation
HPLC: High Performance Liquid Chromatography
JPL: Java Proteomic Library
LC−MS/MS: Mass spectrometry/mass spectrometry, i.e. tandem mass spectrometry
MALDI: matrix-assisted laser desorption
MC: missed cleavages
MGF: Mascot Generic File MRM: Multiple Reaction Monitoring
PITC: phenylisotiocyanate PTC: phenylthiocarbamoyl
ROC: Receiver Operating Characteristic
SSE: Sum of Squared Errors
ST: Semi tryptic

REFERENCES

(1) Eng, J.; McCormack, A.; Yates, J. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.
(2) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20* (18), 3551−67.
(3) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20* (9), 1466−7.
(4) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. *J. Proteome Res.* **2004**, *3* (5), 958−64.
(5) Summerfield, S. G.; Bolgar, M. S.; Gaskell, S. J. *J. Mass Spectrom.* **1997**, *32* (2), 225−231.
(6) van der Rest, G.; He, F.; Emmett, M. R.; Marshall, A. G.; Gaskell, S. J. *J. Am. Soc. Mass Spectrom.* **2001**, *12* (3), 288−95.
(7) Sanchez, A.; Perez-Riverol, Y.; González, L. J.; Noda, J.; Betancourt, L.; Ramos, Y.; Gil, J.; Vera, R.; Padron, G.; Besada, V. *Anal. Chem.* **2010**, *82* (20), 8492−501.
(8) Perez-Riverol, Y.; Sanchez, A.; Ramos, Y.; Schmidt, A.; Muller, M.; Betancourt, L.; Gonzalez, L. J.; Vera, R.; Padron, G.; Besada, V. *J. Proteomics* **2011**, *74* (10), 2071−82.
(9) Wang, D.; Fang, S.; Wohlhueter, R. M. *Anal. Chem.* **2009**, *81* (5), 1893−1900.
(10) Diego, P. A.; Bajrami, B.; Jiang, H.; Shi, Y.; Gascon, J. A.; Yao, X. *Anal. Chem.* **2010**, *82* (1), 23−27.
(11) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. *Mol. Cell. Proteomics* **2010**, *9* (12), 2772−82.
(12) Bern, M.; Goldberg, D.; McDonald, W. H.; Yates, J. R., 3rd *Bioinformatics* **2004**, *20* (Suppl 1), i49−i54.
(13) Michalski, A.; Neuhauser, N.; Cox, J.; Mann, M. *J. Proteome Res.* **2012**, *11* (11), 5479−91.
(14) Sunyaev, S.; Liska, A. J.; Golod, A.; Shevchenko, A. *Anal. Chem.* **2003**, *75* (6), 1307−15.
(15) Tabb, D. L.; Ma, Z. Q.; Martin, D. B.; Ham, A. J.; Chambers, M. C. *J. Proteome Res.* **2008**, *7* (9), 3838−46.
(16) Tabb, D. L.; Saraf, A.; Yates, J. R., III *Anal. Chem.* **2003**, *75* (23), 6415−21.
(17) Baker, S. G. *J. Natl. Cancer Inst.* **2003**, *95* (7), 511−5.
(18) Foettinger, A.; Leitner, A.; Lindner, W. *J. Mass Spectrom.* **2006**, *41* (5), 623−32.
(19) Leng, J.; Wang, H.; Zhang, L.; Zhang, J.; Wang, H.; Cai, T.; Yao, J.; Guo, Y. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (7), 1204−1213.