Published in final edited form as:
Proteomics. 2011 October ; 11(20): 4105–4108. doi:10.1002/pmic.201100297.

Can the false-discovery rate be misleading?

Rodrigo Barboza[1,&], **Daniel Cociorva**[2,&], **Tao Xu**[2], **Valmir C Barbosa**[1], **Jonas Perales**[4], **Richard H Valente**[4], **Felipe M G França**[1], **John R Yates III**[2], and **Paulo C Carvalho**[3,4,*]

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

[2]Department of Chemical Physiology, The Scripps Research Institute, La Jolla, California, USA

[3]Center for Technological Development in Health of the Oswaldo Cruz Foundation, Rio de Janeiro, Brazil

[4]Laboratory of Toxinology, Oswaldo Cruz Institute, Rio de Janeiro, Brazil

Abstract

The decoy-database approach is currently the gold standard for assessing the confidence of identifications in shotgun proteomic experiments. Here we demonstrate that what might appear to be a good result under the decoy-database approach for a given false-discovery rate could be, in fact, the product of overfitting. This problem has been overlooked until now and could lead to obtaining boosted identification numbers whose reliability does not correspond to the expected false-discovery rate. To remedy this, we are introducing a modified version of the method, termed a semi-labeled decoy approach, which enables the statistical determination of an overfitted result.

Keywords

shotgun proteomics; overfitting; protein identification; false-discovery rate; decoy

The decoy-database approach [1,2] is currently the gold standard for assessing identifications in shotgun proteomic experiments. Briefly, this method relies on using a protein identification search engine to match experimental spectra against theoretical ones generated from a database containing target protein sequences and labeled decoys (e.g., reversed target sequences), usually occurring in the same number. According to Elias and Gygi, “one can estimate the total number of false positives (FPs) that meet specific selection criteria by doubling the number of selected decoy hits” [2]. This rationale is applied to the final lists of identifications, which include the decoy hits (i.e., hits known to be incorrect), and in the end the user generally only counts the identifications assigned as target.

The proteomics community has developed tools (e.g., DTASelect [3] and IDPicker [4]) having roots in this approach to automatically filter out low-quality identifications. New filtration tools and methods are usually benchmarked entirely by how many peptide sequence matches (PSMs) are identified under a specified false-discovery rate (FDR) [5]. The ensuing efforts to maximize PSMs tend to push authors to use increasingly complex

[*] **Corresponding author:** Paulo Costa Carvalho, paulo@pcarvalho.com, Laboratório de Toxinologia, Instituto Oswaldo Cruz, Pavilhão Ozório de Almeida, Sala 13, Fundação Oswaldo Cruz-FIOCRUZ, Ave. Brasil, 4365 – Manguinhos, ZIP: 21045-900, Rio de Janeiro, RJ, Brasil., Phone.: (55-21) 2562-1241, Fax : (55-21) Lab. 2562-1410.

[&] Both authors contributed equally.

The authors declared no conflicts of interest.

discriminant functions to improve results under the same FDR. This, in turn, can give rise to a limitation not anticipated in the Elias-Gygi guideline.

Here we show that what might seem to be a good result under the decoy-database approach, henceforth referred to as the labeled decoy approach, can actually be the product of overfitting a discriminant model to the dataset. To overcome this limitation we introduce a modified decoy method, here termed a semi-labeled decoy approach, which relies on labeled decoys, but also on unlabeled decoys; the last one are sequences that the discriminator does not know to be decoys. These unlabeled decoys serve as an internal error reference that helps to statistically deal with overfitting.

We demonstrate the overfitting problem and the effectiveness of our approach on datasets of mass spectra obtained by analyzing *Pyrococcus furiosus* and *Trypanosoma cruzi* lysates with an Orbitrap XL (Thermo, San Jose, CA) under conditions previously described literature [6]. These spectra were searched using ProLuCID [7] against three types of database generated as follows:

1. The first is the widely adopted Target–Reverse database (T-R DB). In it, for every target sequence a decoy sequence is generated by reversing the target. Clearly, this produces a final database with target and decoy sequences in the same number.

2. The second database is here termed the Target-Scrambled0-Scrambled1 database (T-S0-S1 DB). In it, for every target sequence two decoy sequences, S0 and S1, having the same length as the target, are generated by randomly scrambling the contents of the target sequence. The number of non-target sequences is twice that of the target. The reason for generating two layers of decoys is that one will serve as labeled decoys and the other as unlabeled.

3. The third database is referred to as the Target-PairReversed-MiddleReversed database (T-PR-MR DB). In it, for every target sequence a PR and an MR sequence are generated as a function of the digestion enzyme used in the project. For each target sequence, first the peptides that the enzyme in question will produce are listed. For each one a PR peptide is generated by first swapping the two outermost amino acids, then treating pairs of the remaining amino acids as units and reversing their order. To exemplify, given the target peptide ABCDEFGHI, its PR peptide is IGHEFCDBA. The final PR sequence is obtained by concatenating all PR peptides. Similarly, for each target peptide an MR peptide is generated by first swapping the two outermost amino acids, then dividing the remaining portion in half and reversing each of the halves separately. To exemplify, the former target peptide becomes IEDCBHGFA. The final MR sequence is obtained by concatenating all MR peptides.

The T-PR-MR format is useful to reproduce the advantages of the widely adopted target-reverse approach when using two layers of decoys. As is known, randomizing each sequence independently does generate higher peptide diversity than reversing each sequence, especially in proteomes with high redundancy in the sequences or having conserved regions, such as those of mammals. The search engine will then compare each spectrum to more candidates from the randomized sequences than from the target sequences. This will generate a bias when estimating the FDR, as most search engines will not consider the number of distinct peptides generated from each protein database and most FDR computations assume the number of comparisons to targets and decoys to be the same. As an example, the Human IPI database contains some 140% more unique tryptic peptides in the scrambled decoy sequences than in the targets. The T-PR-MR format, similarly to the T-R format, addresses these issues by acceptably reproducing the diversity found when generating decoys. We created the T-PR-MR format by empirically testing different ways of rearranging the sequences to minimize overlapping peaks of theoretically generated mass spectra of corresponding T, PR, and MR peptides. We have included the T-S0-S1 format for benchmarking purposes only, as there are still various groups from the proteomics community that use randomized databases instead of reversed ones. However, for the reasons above and as demonstrated later by our results, we do not recommend using them.

We implemented two widely adopted pattern recognition strategies to generate discriminant models based on the search results to filter out low-quality identifications. These are the well-known Bayesian discriminator and a weightless artificial neural network (WNN), known as WiSARD [8], in its recently improved form [9]. A description of these approaches is available in Supplementary File I. Next we benchmarked both strategies by using the three database formats and accepting a 1% FDR at the spectral level, calculated by dividing the number of PSMs originating from labeled decoys by the total number of PSMs. The results from the *P. furiosus* dataset are presented in Tables I and II.

The results from Tables I and II favor the WNN over the Bayesian discriminator, as the former yielded more PSMs. However, by introducing unlabeled decoys we see that the premise of having roughly the same number of false positives and of decoys does not always withstand detailed scrutiny: the results from the Bayesian discriminator appear to be consistent but those from the WNN do not. We also note that our new DB format (T-PR-MR) yields better results (i.e., less overfitting and more PSMs), as they reflect peptide diversity better than the randomized approach.

Clearly, what lies behind the apparent success of the WNN is related to its increased complexity, which has enabled it to overfit the data (i.e., achieve better separation of the labeled decoys from the rest). In this regard, even though it provided more PSMs meeting the FDR criterion, the elevated number of unlabeled decoys demonstrates that the results are not as reliable as those of the Bayesian discriminator.

As the number of unlabeled decoy identifications is expected to be roughly the same as that of labeled decoys in a database containing target, decoys, and unlabeled decoys in equal numbers, an overfitting p-value can be approximated by $P=\Pr(X>s) \approx \sum_{t=s+1}^{n} \mathrm{Bin}(t, n, p)$. Here X is a random variable indicating the number of unlabeled decoys identified, s is the value of X reported by the discriminator, n is the total number of identifications, p is the expected fraction of unlabeled decoys (i.e., p 0.01 for an FDR of 1%), and Bin is the binomial distribution function. This approach only applies to databases in which peptide diversity can be assumed to be nearly equal among the target, decoy, and unlabeled decoy sequences. By re-analyzing the tables above it follows that indeed only the results from the Bayesian discriminator can be taken with confidence ($P \gg 0.05$) for the experiment at hand.

In a handpicked analysis from the *T. cruzi* dataset we show that the nature of the experiment can lead to overfitting ($P < 0.05$) even for discriminators, like the Bayesian one, that have done well in other circumstances. These results are presented in Table III.

Table III presents a marginally overfitted result for the T-PR-MR approach ($P = 0.03$). Such results could be improved by employing widely adopted *ad-hoc* filtration strategies. Examples are only considering proteins with two spectral counts or two sequence counts. Nevertheless, an effective discriminator still remains the core of a filtration algorithm, as it is the one responsible for sorting the results according to confidence.

From our experience, in general overfitting should not be a problem for the majority of the widely adopted tools (e.g., Scaffold (Proteome Software), DTASelect, IDPicker, etc.), as they have already matured and have been extensively tested by the proteomics community.

Nevertheless, it can be advisable to test for overfitting when experimenting with new parameters, even for a tool that one has experience with. Once it has been verified that overfitting is not a problem, it becomes unnecessary to search in databases with two layers of decoys, with the advantages of avoiding the increased search time and the loss in sensitivity caused by a database with more decoys and therefore more “distractions” for the search engine.

We strongly recommend adopting the semi-labeled decoy approach when benchmarking new tools. Without proper awareness, it would be easy to advocate in favor of our WNN discriminator over the Bayesian one. In fact, we claim we can ultimately build a filtration tool capable of outperforming any of the widely adopted filtration tools under current benchmarking standards (i.e., number of PSMs under a given FDR). As we demonstrated, however, its results would not be trustworthy.

In summary, the semi-labeled decoy approach complements the labeled decoy approach by statistically dealing with the overfitting problem. The method is simple and therefore makes it easy for authors of filtration tools to adopt overfitting p-values. As far as we know, this is the first strategy that can empirically demonstrate the overfitting in proteomic FDR experiments. We have limited ourselves to demonstrating the approach at the spectral level, but variations can be easily developed for use at the peptide and protein levels. Most importantly, we have shown that basing one’s decision exclusively on the FDR can be misleading. The mass spectra, search databases, and Java source code generated for this study are available at: http://max.ioc.fiocruz.br/pcarvalho/overfitting/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support provided by: CAPES-Fiocruz 30/2006, PDTIS-Fiocruz, CNPq 306070/2007-3, FAPERJ-BBP, NIH P41 RR011823 and ROI MH067880.

Reference List

1. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res. 2003; 2:43–50. [PubMed: 12643542]
2. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007; 4:207–214. [PubMed: 17327847]
3. Cociorva D, Tabb L, Yates JR. Validation of tandem mass spectrometry database search results using DTASelect. Curr Protoc Bioinformatics. 2007 Chapter 13: Unit 13.4.
4. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res. 2009; 8:3872–3881. [PubMed: 19522537]
5. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007; 4:923–925. [PubMed: 17952086]
6. Washburn MP, Wolters D, Yates JR III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol. 2001; 19:242–247. [PubMed: 11231557]
7. Xu T, Venable JD, Park SK, Cociorva D, Lu B, Liao L, et al. ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. Mol Cell Proteomics. 2006; 5 S:174.

8. Aleksander I, Thomas W, Bowden P. WiSARD, a radical new step forward in image recognition. Sensor Rev. 1984; 4:120–124.

9. Grieco BPA, Lima PMV, De Gregorio M, França FMG. Producing pattern examples from "mental" images. Neurocomputing. 2010; 73:1057–1064.

Table I

Bayesian discriminator results on the *P. furiosus* dataset

T-R DB, T-S1-S0 DB, and T-PR-MR DB are the databases described in the main text. The numbers in the Spectra rows indicate how many PSMs were obtained. The numbers in the Peptides rows indicate how many unique peptides were identified.

<table>
<tr><th rowspan="2"></th><th colspan="4">T-R DB</th></tr>
<tr><th>Labeled Decoys</th><th colspan="2">Total Target</th><th>Total</th></tr>
<tr><td>Spectra</td><td>1073</td><td colspan="2">106280</td><td>107353</td></tr>
<tr><td>Peptides</td><td>900</td><td colspan="2">19532</td><td>20432</td></tr>
<tr><td></td><td colspan="4">T-S1-S0 DB</td></tr>
<tr><td></td><td>Labeled Decoys</td><td>Unlabeled Decoys</td><td>Total Target</td><td>Total</td></tr>
<tr><td>Spectra</td><td>1083</td><td>1105</td><td>106184</td><td>108372</td></tr>
<tr><td>Peptides</td><td>982</td><td>983</td><td>19609</td><td>21574</td></tr>
<tr><td></td><td colspan="4">T-PR-MR DB</td></tr>
<tr><td>Spectra</td><td>1083</td><td>1064</td><td>106229</td><td>108376</td></tr>
<tr><td>Peptides</td><td>936</td><td>939</td><td>19587</td><td>21462</td></tr>
</table>

Table II
WNN discriminator results on the *P. furiosus* dataset

T-R DB, T-S1-S0 DB, and T-PR-MR DB are the databases described in the main text. The numbers in the Spectra rows indicate how many PSMs were obtained. The numbers in the Peptides rows indicate how many unique peptides were identified.

	T-R DB			
	Labeled Decoys	**Total Target**		**Total**
Spectra	1162	115074		116236
Peptides	1099	26142		27241
	T-S1-S0 DB			
	Labeled Decoys	Unlabeled Decoys	Total Target	Total
Spectra	1150	4917	108945	115012
Peptides	1126	4513	22714	28353
	T-PR-MR DB			
Spectra	1152	4656	109440	115248
Peptides	1100	4291	22803	28194

Table III
Bayesian discriminator results on the *T. cruzi* dataset

T-R DB, T-S1-S0 DB, and T-PR-MR DB are the databases described in the main text. The numbers in the Spectra rows indicate how many PSMs were obtained. The numbers in the Peptides rows indicate how many unique peptides were identified.

	T-S1-S0 DB			
	Labeled Decoys	**Unlabeled Decoys**	**Total Target**	**Total**
Spectra	12	43	1221	1276
Peptides	9	12	267	288
	T-PR-MR DB			
Spectra	12	20	1235	1267
Peptides	11	17	273	301