A Complex Standard for Protein Identification, Designed by Evolution

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · APRIL 2012

Impact Factor: 4.25 · DOI: 10.1021/pr300055q · Source: PubMed

6 AUTHORS, INCLUDING:

Julia Maria Burkhart
Leibniz-Institut für Analytische Wissenschaften
20 PUBLICATIONS **453** CITATIONS
SEE PROFILE

René Zahedi
Leibniz-Institut für Analytische Wissenschaften
81 PUBLICATIONS **2,693** CITATIONS
SEE PROFILE

Albert Sickmann
Leibniz-Institut für Analytische Wissenschaften
233 PUBLICATIONS **8,249** CITATIONS
SEE PROFILE

Lennart Martens
Ghent University
211 PUBLICATIONS **6,559** CITATIONS
SEE PROFILE

Technical Note

A complex standard for protein identification, designed by evolution

Marc Vaudel[1,$], Julia M. Burkhart[1,$], Daniela Breiter[1,2], René P. Zahedi[1], Albert Sickmann[2,3] and Lennart Martens[4,5]*

[1] Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Dortmund, Germany

[2] Department of Statistics, Dortmund University of Technology, 44221 Dortmund, Germany

[3] Medizinisches Proteom-Center (MPC), Ruhr-Universität, Bochum, Germany

[4] Department of Medical Protein Research, VIB, Ghent, Belgium

[5] Department of Biochemistry, Ghent University, Ghent, Belgium

[$] These authors contributed equally to the work

* Corresponding author:

Prof. Dr. Albert Sickmann
tel: +49 231 1392 100
fax: +49 231 1392 200
email: albert.sickmann@isas.de

ABSTRACT

Shotgun proteomic investigations rely on the algorithmic assignment of mass spectra to peptides. The quality of these matches is therefore a cornerstone in the analysis and has been the subject of numerous recent developments. In order to establish the benefits of novel algorithms, they are applied to reference samples of known content. However, these were recently shown to be either too simple to resemble typical real-life samples, or as leading to results of lower accuracy as the method itself.

Here, we describe how to use the proteome of *Pyrococcus furiosus*, a hyperthermophile, as a standard to evaluate proteomics identification workflows. Indeed, we prove that the *Pyrococcus furiosus* proteome provides a valid method for detecting random hits, comparable to the decoy databases currently in popular use, but we also prove that the *Pyrococcus furiosus* proteome goes squarely beyond the decoy approach by also providing many hundreds of highly reliable true positive hits. Searching the *Pyrococcus furiosus* proteome can thus be used as a unique test that provides the ability to reliably detect both false positives as well as proteome-scale true positives, allowing the rigorous testing of identification algorithms at the peptide and protein level.

KEYWORDS: Peptide identification, protein identification, Bioinformatics, Pyrococcus furiosus

INTRODUCTION

In shotgun proteomics, vast numbers of fragmentation mass spectra are associated to peptides that are in turn used to infer proteins. Since the conclusions of such investigations rely heavily on this crucial step of the workflow, substantial efforts have been dedicated to maximizing the yield of identified spectra while simultaneously controlling the quality of the results. One of the major breakthroughs in the field was the introduction of Target/Decoy database searches to enable quality control of the identification process through the estimation of False Discovery Rates (FDR)[1]. Various algorithms[2-5] were subsequently developed in order to maximize the count of identifications at a defined, estimated FDR.

It is obviously important to test the actual performance of these novel methods, a process that can be challenging as actual FDRs to benchmark estimated FDRs are very difficult to obtain. Typically, this problem is addressed by using standard samples of known content[6], since knowledge of the sample composition enables the actual discrimination between false and true positive assignments and hence the evaluation of the method's efficiency. For instance, spectra obtained from the Sigma UPS1 standard (comprising 48 human proteins and minor contaminants) searched against the human complement of the UniProtKB/Swiss-Prot database (20,260 target sequences, 4th of November 2010) ensure with a very high confidence that peptide to spectrum matches (PSMs) pointing to UPS1 proteins are correct matches whereas all other matches can be considered as incorrect[7]. However, such a small set of proteins is hardly representative of the complexity of samples typically encountered; Indeed, it is obvious that a reliable 1% FDR at the protein level cannot be reached with only 48 true positive proteins. Moreover, recent studies demonstrated the lack of precision of other available standards[8].

There is hence a need from the community for a standard providing both protein complexity and known content. In the present work, we describe the design of a representative reference identification workflow, and how its reliability can be verified. We demonstrate that the *Pyrococcus furiosus* (Pfu) proteomic standard makes a suitable reference, providing a sufficient number of true positive proteins

while ensuring that false positive matches can be accurately discriminated. Moreover, we show that standards at different evolutionary distances yield various levels of accuracy when estimating reference identification metrics.

MATERIAL AND METHODS

Material

Ammonium hydrogen carbonate (NH_4HCO_3), guanidinium hydrochloride (Gu-HCl), iodoacetamide (IAA) and trypsin, triflouroethanol (TFE) were purchased from Sigma-Aldrich, Steinheim, Germany. Sodium di-hydrogen phosphate (NaH_2PO_4) were purchased from Merck KGaA, Darmstadt, Germany. Trichloroacetic acid (TCA) was obtained from Roth, Karlsruhe, Germany. DTT was bought from Roche Diagnostics, Mannheim, Germany. Bicichinon assay (BCA) was acquired from Pierce Thermo Fisher Scientific, Schwerte, Germany and Spec C18AR tips as well as Complex Proteomics Standard (representing the proteome of Pyrococcus furiosus, Pfu) from Agilent Technologies, Darmstadt, Germany. All chemicals for ultrapure HPLC solvents such as formic acid (FA), trifluoro acetic acid (TFA) and acetonitrile (ACN) were obtained from Biosolve, Valkenswaard, the Netherlands.

Pyrococcus furiosus sample

Pyrococcus furiosus (Pfu) was treated according to the manufacturer's instructions. Briefly, an aliquot of 100 µg dissolved in 2 M GuHCl 50 mM Na_2HPO_4 was precipitated with TCA and subsequently dissolved in 50 mM NH_4HCO_3, 4 mM DTT , and 50% TFE. Disulfide bonds were reduced for 60 min at 56°C and afterwards free sulfhydryl groups were carbamidomethylated using 15 mM IAA for 60 min at room temperature in the dark. For digestion TFE was reduced to final concentration of 5% with 50 mM NH_4HCO_3 and trypsin was added in a protease to protein ratio of 1:30 and incubated at 37°C overnight.

Digests were controlled using monolithic column separation (PepSwift monolithic PS-DVB PL-

CAP200-PM, Dionex) on an inert Ultimate 3000 HPLC (Dionex, Germering, Germany) by direct injection of 0.1 µg sample and a binary gradient (solvent A: 0.1 % TFA, solvent B: 0.08 % TFA, 84 % ACN) with a flow rate of 2.2 µl/min at 60°C. UV traces were acquired at 214 nm. Although the efficiency of tryptic digest of *Pyrococcus furiosus* is controversial[9], it is clear from supplementary Figure 1 that the digestion efficiency was sufficient here, a finding that is also indirectly confirmed by the large variety of proteins identified.

MS analysis

Nano-LC-MS/MS was performed on an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) coupled to an Ultimate 3000 Rapid Separation Liquid Chromatography (RSLC) system (Dionex, Germering, Germany). Briefly, peptides were preconcentrated on a 100 µm ID reversed-phase (RP) trapping column (Acclaim PepMap RSLC 100 µm x 2 cm, 3 µm particle size, 100 Å pore size, Dionex) in 0.1% TFA followed by separation on a 75 µm ID RP column (Acclaim PepMap RSLC 75 µm x 25 cm, 2 µm particle size, 100 Å pore size, Dionex) using a binary gradient (solvent A: 0.1% FA and solvent B: 0.1% FA 84% ACN) ranging from 5-50% of solvent B at a flow rate of 300 nL/min in 90 min. MS survey scans were acquired in the range of 300 to 2,000 m/z at a resolution of 30,000 using the polysiloxane at m/z 371.101236 as lock mass[10]. The ten most intensive signals were subjected to HCD-MS/MS taking into account a dynamic exclusion of 10 s. HCD spectra were acquired with a normalized CE of 35%, a precursor isolation width of 2.0 m/z and an activation time of 0.1 ms with a resolution of 7,500. Orbitrap AGC target values were set to 10^6 for MS and $2*10^5$ for MS^n.

Spectrum identification

Raw data were converted into mzML[11] files using msconvert as part of the Proteowizard 1.6.0 package[12]. They were further converted into mgf files using OpenMS 1.8[13]. Database searches with OMSSA[14] (version 2.1.9) and X!Tandem[15] (version 2010.12.01.1) were conducted with the help of

SearchGUI[16] (version 1.6). Database searches with Mascot[17] (version 2.3) were conducted via Mascot Daemon.

Search settings were: a maximum of one allowed missed cleavage, peptide charges 2+ – 4+, peptide mass tolerance of 10 ppm, fragment ion mass tolerance of 0.5 Da, carbamidomethylation of Cys as fixed, and both phosphorylation of Ser/Thr/Tyr as well as oxidation of Met as variable modifications. All other settings were kept at the default values of SearchGUI.

Spectra were searched against various complements of the UniProtKB/Swiss-Prot database[18]: (1) all cellular organisms (downloaded on the 4th of August 2011, 15,813,946 target sequences), (2) *Archaea* (downloaded on the 25th of July 2011, 295,106 target sequences), (3) *Thermococci* (downloaded the 3rd of August 2011, 21,615 target sequences), (4) *Pyrococcus* (downloaded on the 3rd of August 2011, 8,408 target sequences), (5) *Pyrococcus furiosus* (downloaded on the 3rd of August 2011, 2,139 target sequences), (6) *Eukaryota* (downloaded on the 25th of July 2011, 4,465,416 target sequences), (7) *Vertebrata* (downloaded on the 25th of July 2011, 850,859 target sequences), (8) *Mammalia* (downloaded on the 25th of July 2011, 443,006 target sequences), (9) *Homo sapiens* (downloaded on the 4th of November 2010, 20,260 target sequences). Moreover, the following databases were generated by manipulating the fasta files[19]: (1 without Pfu) cellular organisms without Pfu sequences, (2 without Pfu) *Archaea* without Pfu sequences, (3 without Pfu) *Thermococci* without Pfu sequences, (4 without Pfu) *Pyrococcus* without Pfu sequences, (6 with Pfu) *Eukaryota* with Pfu sequences, (7 with Pfu) *Vertebrata* with Pfu sequences, (8 with Pfu) *Mammalia* with Pfu sequences and (9 with Pfu) *Homo sapiens* with Pfu sequences.

For each of the abovementioned databases, a concatenated Target/Decoy version was generated using SearchGUI by reversing the target sequences to obtain the decoy sequences. Generally, spectra matching peptide sequences that were shared between target and decoy databases were omitted. Peptides with less than eight or more than twenty amino acids were not taken into account. Furthermore,

PSMs with an OMSSA e-value higher than 10 were also not considered. Searching the 15,373 MS/MS spectra acquired from 1.5 µg of Pfu against human retrieved 9 PSMs before the first Decoy hit (OMSSA e-value of $6.58*10^{-6}$). Seven of these matched to keratin, one to O60675 (OMSSA e-value of $4.48*10^{-6}$) and one to either Q00610 or P53675 (OMSSA e-value of $2.39*10^{-6}$). These hits are thus either due to contamination, either a memory effect in the column[20] or false positive identifications as suggested by the similarity of the e-values of the two later hits and the e-value of the first decoy hit. The corresponding 9 spectra were thus removed from the spectrum list in order to ensure the quality of the study.

FDR estimation

At a defined threshold α, the number of retained false positive PSMs (n_{FP}) divided by the total number of retained PSMs (n_{PSM}) represents the False Discovery Rate (FDR)[21]:

$$FDR(\alpha) = \frac{n_{FP}}{n_{PSM}} \quad (1)$$

Generally, the FDR is estimated by searching the spectra against a concatenated Target/Decoy database with equal amounts of Target and Decoy sequences where a random match has equal chances to hit both databases[1] – there is thus no need for a correction factor. The number of decoy hits is hence an estimator for the quantity of (target) false positives (n_{FP} in equation 1). When sorting the PSMs according to the OMSSA e-value, it is thus possible to estimate the FDR at any e-value α ($\hat{FDR}(\alpha)$) by dividing the number of decoy hits with an e-value smaller than α (n_{decoy}) by the number of target hits with an e-value smaller than α (n_{target}):

$$\hat{FDR}(\alpha) = \frac{n_{decoy}}{n_{target}} \quad (2)$$

Searching the *Pyrococcus furiosus* MS/MS spectra against the *Pyrococcus furiosus* database allowed the validation of 8391 PSMs at 1% FDR, representing 4726 different peptide sequences derived from

839 proteins.

Definition of a false positive match

A false positive match between a spectrum and a peptide is here strictly understood as a random match. Close but wrong sequences as well as post-translational modification localization errors are not taken into account as these are not assessable by target/decoy strategies[22,23] nor by complex standards of known content.

RESULTS

Design of standard identification procedures

To demonstrate that a protein identification method is reliable, it is commonly tested on proteomic standards of known content. These analyses allow the discrimination of true and false positives based on the known sample composition, and thus allow the estimation of reference metrics like a reference False Discovery Rate (FDR). The PSMs pointing to the proteins actually in the sample (the UPS1 proteins in the example given in the introduction) are considered as true positive matches whereas all other hits are suspected to be errors. As described by Granholm et al.[8] and illustrated in Figure 1, the (target) database can be subdivided into two parts: (A) the sequences of proteins that can be expected to be in the sample, and (B) the entrapment sequences that are only hit by false positives. The corresponding decoy sequences are designated (A') and (B').

The necessary and sufficient conditions for the accurate establishment of reference metrics are then: (1) no true positive PSM shall hit the entrapment database, and (2) no random match shall hit the sample sequences. Meeting (1) is very difficult due to the presence of contaminants: in the case of the UPS1 example, contaminants like Ig antibodies will generate confident hits in the entrapment database which can be erroneously flagged as false positives. (2) depends on the respective sizes of the sample and entrapment databases: in the case of the UPS1 example, the number of peptides that can be derived from UPS1 sequences is small (486 without accounting variable modifications) compared to the number of human peptides (346,367 without accounting variable modifications), so the probability for a random match to hit the sample database is negligible ($1.4*10^{-3}$).

These conditions are hence not met by all standards and their verification is often lacking in the literature. In order to address this issue, we provide here straightforward experimental methods to verify that these conditions are met: in order to verify (1), a search can be performed merely against the entrapment database; here, no confident hit should be found. As mentioned in the Materials and

Methods section, spectra are searched against a concatenated Target/Decoy database. A random match thus has the same chance to randomly occur within both, the sample sequences (A in Figure 1) and the decoy versions of the sample sequences (A' in Figure 1). By design however, no true positive match hits the decoy sequences. Thus, (2) can be readily verified by counting the matches that hit the decoy versions of the sample sequences[1,7] (A' in Figure 1); this number should be equal to zero to satisfy (2). It is important to note here that condition (2) has not been verified in previously published studies[8].

A standard of choice: *Pyrococcus furiosus*

Pyrococcus furiosus (Pfu) is an extremophile well known for its fast replication cycle and for its unique notion of comfort: its optimum growth temperature is 100°C[24]. It has therefore been the subject of many biological studies[25-29]. Moreover, as suggested by its phenotype, its genome is dramatically different from most other species. Indeed, only seven tryptic peptides can be found in common between the human and Pfu protein databases (respectively 346,367 and 20,710 tryptic peptides). However, despite this large difference in actual sequences, an analysis of the tryptic peptides shows that these proteomes actually present very similar overall features, providing similar behaviour in LC/MS systems (see supplementary Figure 2).

Pfu is thus a good potential candidate for a standard, since the acquired fragmentation spectra are unlikely to match to peptide sequences derived from commonly studied species. Moreover, the size of its database (2,091 target sequences) suggests that a sufficient quantity of proteins can be identified to obtain a statistically significant size for the protein result set.

The impact of evolutionary distance on the suitability of the entrapment database

Evolution disambiguation between Pfu and human occurred very early in the currently established evolution tree: divergence is estimated at more than a billion years ago, and the effects of traversing this evolutionary distance in the search space used, are shown in Figure 2. A comparison of the shared

tryptic peptides between Pfu and the *Eukaryota*, *Vertebrata*, *Mammalia*, and *Homo sapiens* databases (respectively 151, 16, 9 and 7 shared peptides; red bars in Figure 2) demonstrates that larger evolutionary distances yield increasingly clearly differentiated proteomes. In contrast, the *Archaea*, *Thermococci* and *Pyrococcus* databases without Pfu sequences represented 6,442, 6,350 and 5,196 tryptic peptides shared with the Pfu database, respectively. Pfu sequences are thus clearly differentiated from *Eukaryota* species, and this enables the validation of condition (1): searching our 15,365 Pfu MS/MS spectra against the *Eukaryota*, *Mammalia* and *Homo sapiens* databases did not retrieve any confident hits (excluding the handful of peptides that are shared with Pfu, detailed above). When searching *Vertebrata* sequences, only two PSMs were found (both for the same peptide with sequence SPMGLLLEALGQQEEK) before the first decoy PSM was encountered. In contrast, 2,647, 3,880 and 3,409 PSMs could be validated at 1% FDR when searching the *Archaea*, *Thermococci* and *Pyrococcus* databases without Pfu sequences, respectively (orange bars in Figure 2). The striking difference between these two series of results clearly demonstrates how the selection of the entrapment databases has a crucial impact on the validation of condition (1). Here, it is clear that the whole *Eukaryota* branch validates this condition and the further we extend the evolutionary distance towards the human branch, the better this condition is fulfilled.

On the other hand, the smaller the entrapment database the higher the probability for a random match to hit the sample database (A in Figure 1). Indeed, among our 15,365 Pfu MS/MS spectra, 0, 3, 6, and 16 spectra matched the decoy versions of the Pfu sequences, indicating that a similar number of random matches may have hit the sample database[1,7], when searching against *Eukaryota*, *Vertebrata*, *Mammalia* and *Homo sapiens* databases with addition of the Pfu sequences (blue bars in Figure 2). It is clear however, that this low set of hits (maximum 0.1% of spectra) does not impair the estimation of reference metrics. Additionally, it is obvious that condition (2) is met more readily early in the *Eukaryota* branch, as these earlier databases offer many more entrapment sequences.

Experimental validation

In order to prove experimentally the use of the Pfu standard as a means to validate identification strategies, we verified that the number of entrapment hits indeed corresponded to the number of false positives. In Figure 3A the number of entrapment hits for Pfu spectra is plotted when using *Eukaryota* sequences as entrapment database against the number of decoy hits. It is clear that the number of entrapment hits clearly reflects the quantity of target false positives (estimated using the decoy hits). Moreover, in this search, 3907 different peptide sequences belonging to 777 Pfu proteins could be identified using OMSSA. This approach thus allows the accurate discrimination of true and false positives, and crucially, provides for the first time several hundreds of proteins to validate protein-level identification strategies.

Figure 3A demonstrates a clear agreement between the number of entrapment hits and decoy hits for OMSSA and Mascot, while X!Tandem introduced approximately one hundred additional entrapment hits, illustrating the well known problem of multi-stage search strategies – i.e. searches validating matches based on *a priori* good hits – not being compatible with the Target/Decoy strategy[22]. Interestingly, while benchmarking algorithms based on decoy hits alone, X!Tandem clearly outperforms both other algorithms (see Figure 3B), this is no longer the case when looking at entrapment hits instead (see Figure 3C), where the three algorithms perform similarly – although the number of false positives introduced by the second pass search might still be underestimated.

DISCUSSION

In the present work, we described the two necessary and sufficient conditions for the use of standard samples of known content to evaluate identification methods in mass spectrometry based proteomics. We provided two experimentally straightforward tests that can be applied to verify whether the conditions are validated. Applying this rigorous approach, we could demonstrate that *Pyrococcus furiosus* represents an excellent standard when coupled with the *Eukaryota*, *Vertebrata*, *Mammalia* and *Homo sapiens* sequences as entrapment databases. This standard does not only provide outstanding performance, it furthermore provides sufficient potentially identifiable proteins to robustly validate entire proteomic workflows. An experimental validation was given by the agreement between decoy hits and entrapment hits. While the agreement was excellent with Mascot and OMSSA, it is important to note that multi-stage search procedures (as illustrated here with X!Tandem) are not adapted to such strategies as already discussed in the literature [22]. Moreover, the benefit of the second pass search in terms of the number of true positives is not clear in the present experiment.

Although it is tempting to systematically use the *Eukaryota* database as entrapment database, since it allows the lowest number of random matches in the sample database while avoiding true matches in the entrapment database, it is important to stress that the size of the database tremendously increases the search time: for practical reasons smaller databases might be used instead. Also, using larger databases will provide more opportunities for false positive identifications. Indeed, as illustrated by the green bars in Figure 2, the set of PSMs validated at 1% FDR decreased from 7,798 when searching against the human database with addition of Pfu sequences, to 7,222, 6,751 and 5,002 when searching against the *Mammalia*, *Vertebrata* and *Eukaryota* databases with Pfu sequences, respectively. This result also highlights how crucial it is to tailor the database for species actually expected in the sample. Moreover, according to the general trend observed here, doubling the size of the human database with addition of Pfu sequences would result in a loss of 0.36% PSMs. This value clearly illustrates that the loss of

identifications introduced by the adjunction of decoy sequences in concatenated Target/Decoy approaches is negligible.

Generally, a large set of correct PSM identifications is required for the sake of statistical significance of the investigations at the protein level. It is thus necessary to balance between the quality of the reference metrics and the quantity of correctly assigned matches. Here, we demonstrated that all four databases provide sufficient performance to thoroughly evaluate identification workflows while still providing hundreds of (correct) protein identifications.

ACKNOWLEDGMENTS

The financial support by the Ministerium für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen and by the Bundesministerium für Bildung und Forschung (SARA, DYNAMO) is gratefully acknowledged. L.M. acknowledges the financial support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks") and the PRIME-XS project funded by the European Union 7th Framework Program under grant agreement number 262067.

Supporting Information Available

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

FIGURE CAPTIONS

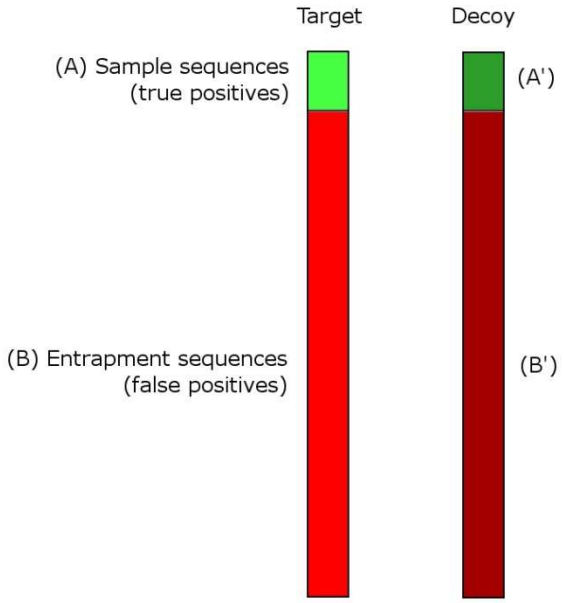


Figure 1. When using a sample of known content for the validation of identification strategies, the search is conducted against a database containing the expected sample sequences (A) as well as entrapment sequences (B). Every sequence has its reversed equivalent in the decoy database (A') and (B').

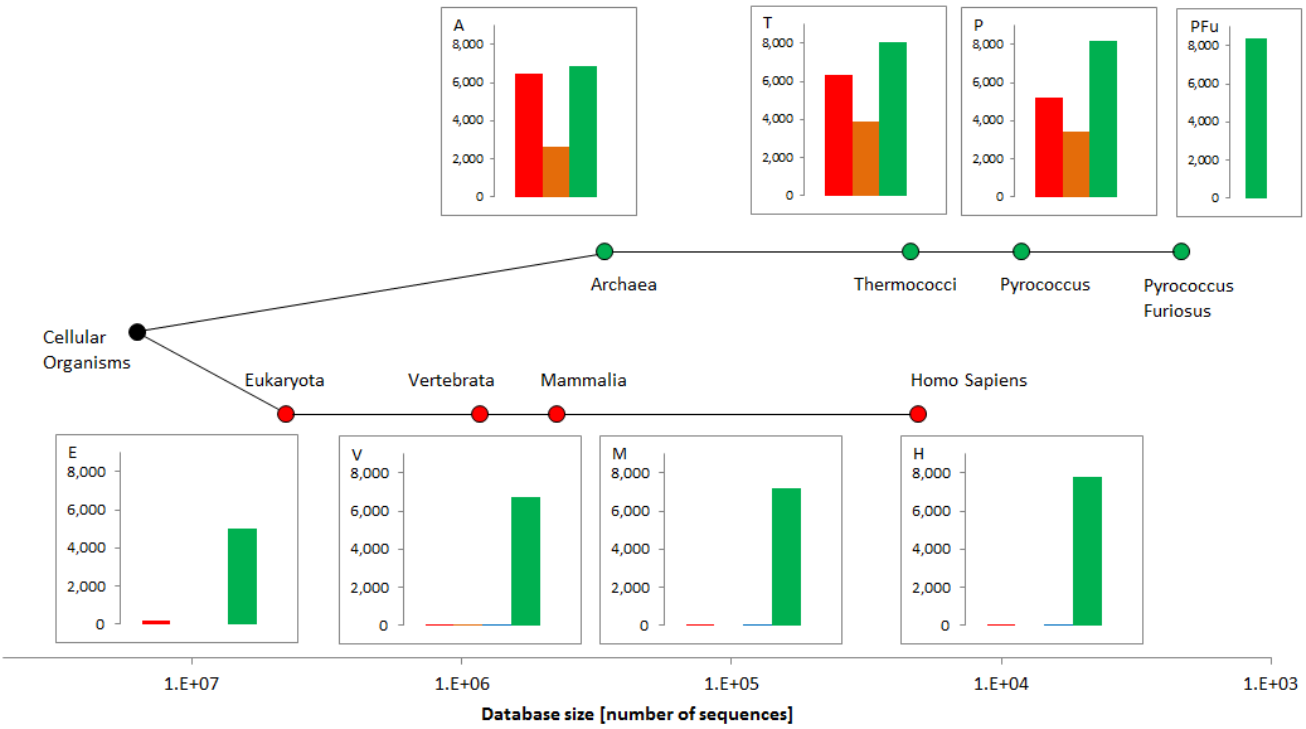
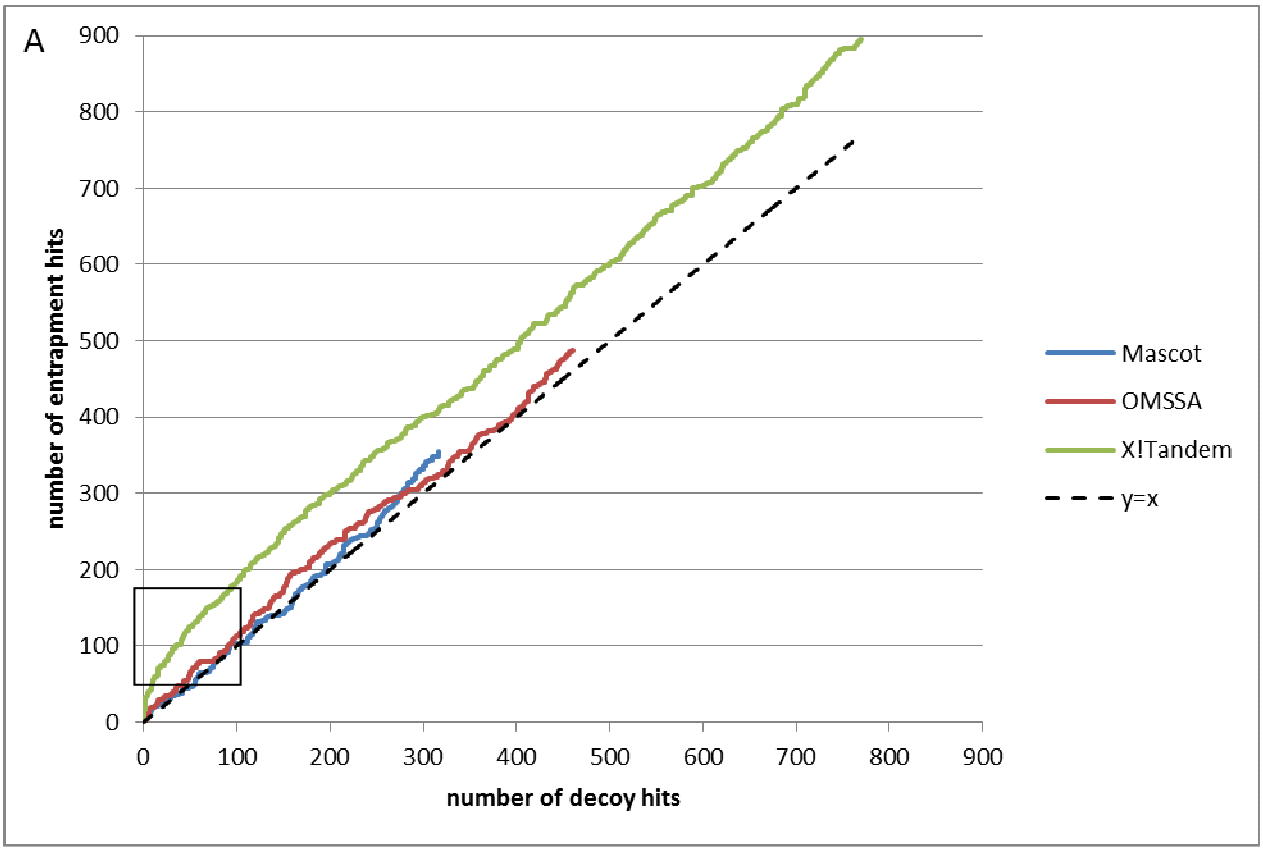
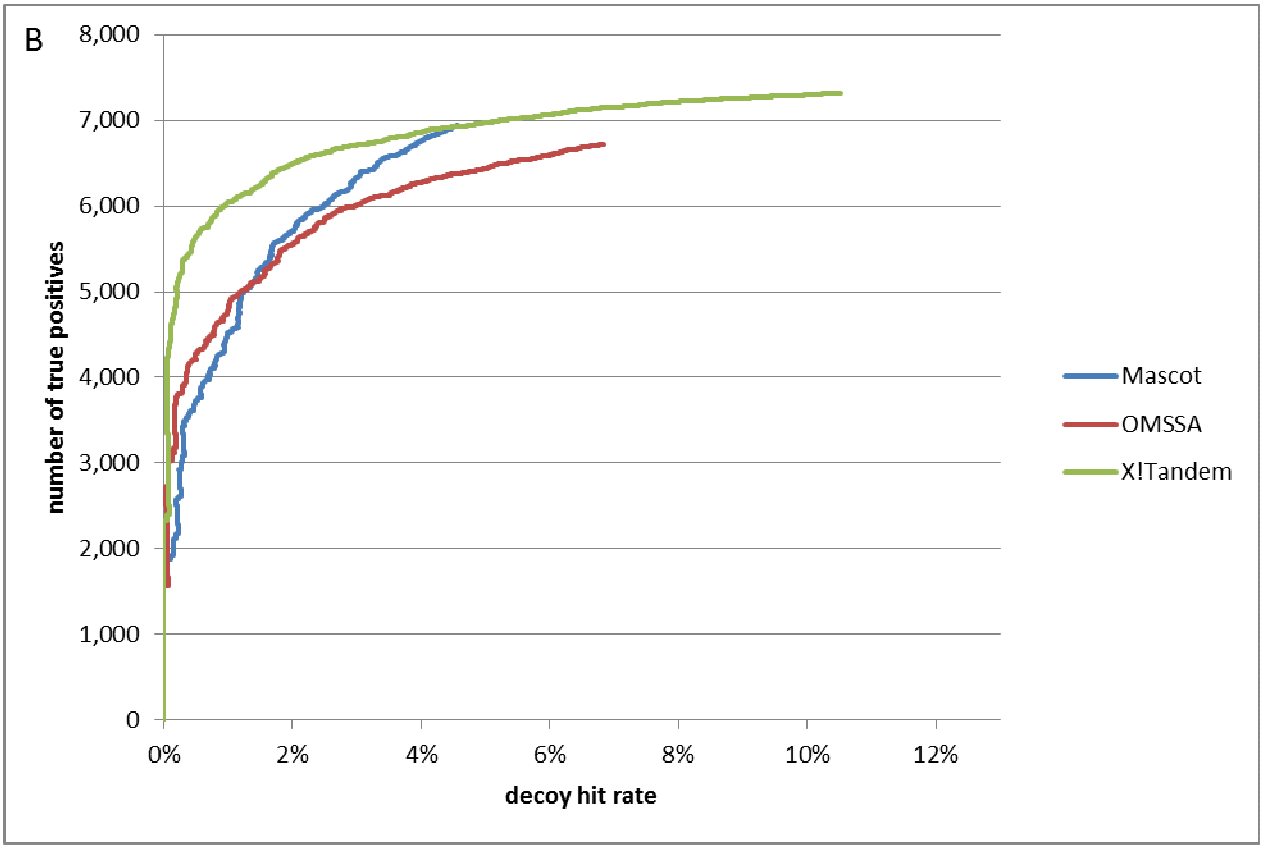


Figure 2. The disambiguation between *Pyrococcus furiosus* (Pfu) and humans occurred very early in evolution. Consequently, *Eukaryota* (E), *Vertebrata* (V), *Mammalia* (M) and *Homo sapiens* (H) databases can be used as entrapment sequences. This is demonstrated by the negligible count (1) of shared tryptic peptides (red), (2) of validated PSMs when searches are performed merely against the entrapment sequences (orange), and (3) of hits in the decoy versions of the sample sequences (A' in Figure 1) (blue), when compared to *Archaea* (A), *Thermococci* (T) and *Pyrococcus* (P). Note that hits in the decoy versions of the sample sequences (blue) cannot be estimated for *Archaea* (A), *Thermococci* (T) and *Pyrococcus* (P) as these databases contain Pfu sequences. Finally, the number of Pfu hits validated at 1% FDR (green) substantially decreases with growing database size.


A
900
800
700
600
500
400
300
200
100
0
number of entrapment hits
0
100
200
300
400
500
600
700
800
900
number of decoy hits
Mascot
OMSSA
X!Tandem
y=x



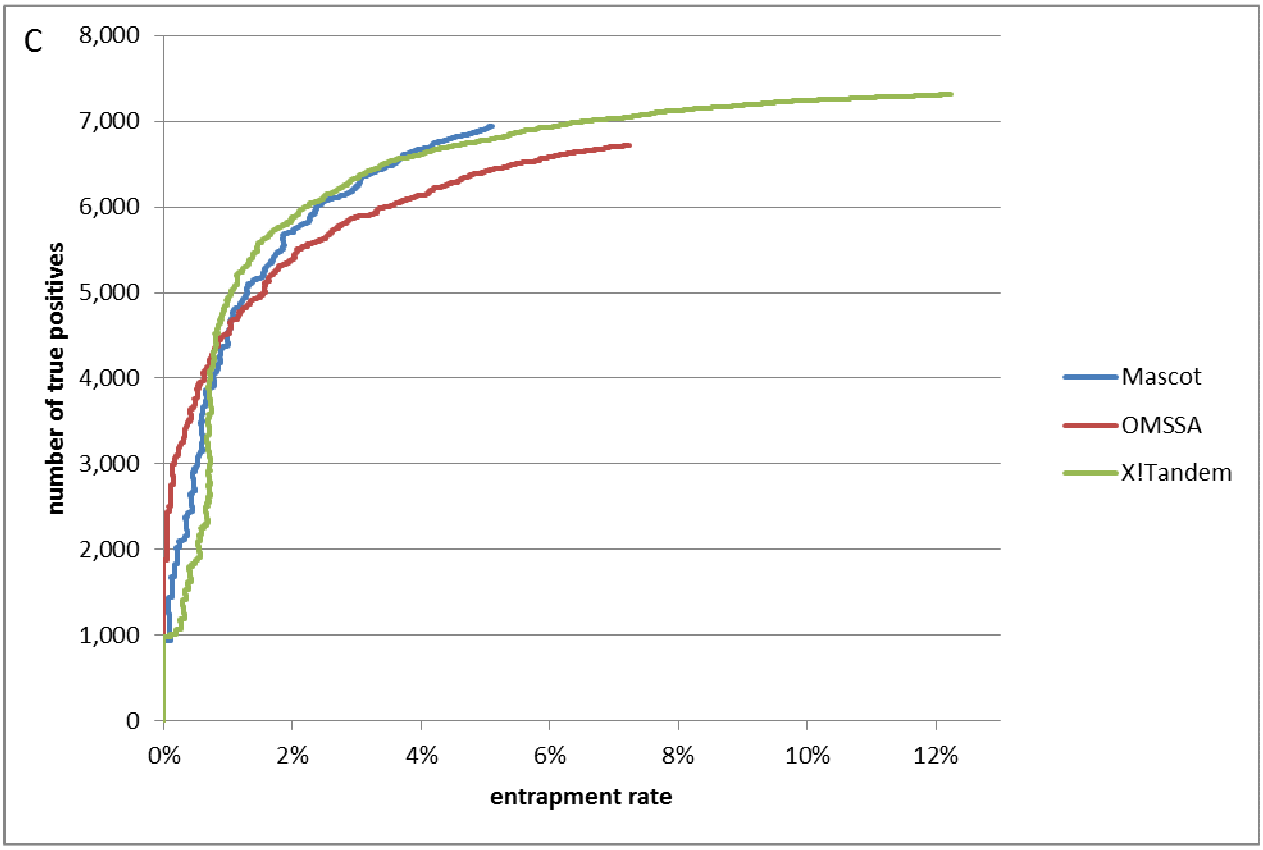


Figure 3. (A) When searching MS/MS spectra obtained from the *Pyrococcus friosus* standard using

Eukaryota sequences as entrapment database, the number of PSMs hitting the entrapment database accurately follows the number of decoy hits. Since decoy hits are an estimator of the quantity of random hits, this agreement demonstrates the efficiency of the method. Similar performance is obtained with OMSSA and Mascot. However, due to its built-in second pass search, X!Tandem generates more entrapments hits than decoy hits. (B) The Reciever Operator Characteristic (ROC) curve based on the decoy hits would indicate that X!Tandem is overperforming when compared to OMSSA and Mascot. (C) However, the same curve based on entrapment hits actually shows that the three search engines demonstrate a very similar performance.

REFERENCES

(1) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007**, *4*, 207.

(2) Kall, L.; Storey, J. D.; Noble, W. S. QVALITY: non-parametric estimation of q-values and posterior error probabilities. *Bioinformatics* **2009**, *25*, 964.

(3) Nahnsen, S.; Bertsch, A.; Rahnenfuhrer, J.; Nordheim, A.; Kohlbacher, O. Probabilistic Consensus Scoring Improves Tandem Mass Spectrometry Peptide Identification. *J Proteome Res* **2011**.

(4) 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* **2006**, *24*, 1285.

(5) Yu, W.; Taylor, J. A.; Davis, M. T.; Bonilla, L. E.; Lee, K. A.; Auger, P. L.; Farnsworth, C. C.; Welcher, A. A.; Patterson, S. D. Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines. *Proteomics* **2010**, *10*, 1172.

(6) Klimek, J.; Eddes, J. S.; Hohmann, L.; Jackson, J.; Peterson, A.; Letarte, S.; Gafken, P. R.; Katz, J. E.; Mallick, P.; Lee, H.; Schmidt, A.; Ossola, R.; Eng, J. K.; Aebersold, R.; Martin, D. B. The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. *J Proteome Res* **2008**, *7*, 96.

(7) Vaudel, M.; Burkhart, J. M.; Sickmann, A.; Martens, L.; Zahedi, R. P. Peptide identification quality control. *Proteomics* **2011**, *11*, 2105.

(8) Granholm, V.; Noble, W. S.; Kall, L. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *J Proteome Res* **2011**, *10*, 2671.

(9) Lee, A. M.; Sevinsky, J. R.; Bundy, J. L.; Grunden, A. M.; Stephenson, J. L., Jr. Proteomics of Pyrococcus furiosus, a hyperthermophilic archaeon refractory to traditional methods. *J Proteome Res* **2009**, *8*, 3844.

(10) Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* **2005**, *4*, 2010.

(11) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Rompp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P. A.; Deutsch, E. W. mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* **2011**, *10*, R110 000133.

(12) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534.

(13) Bertsch, A.; Gropl, C.; Reinert, K.; Kohlbacher, O. OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol Biol* **2011**, *696*, 353.

(14) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J Proteome Res* **2004**, *3*, 958.

(15) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466.

(16) Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011**, *11*, 996.

(17) 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* **1999**, *20*, 3551.

(18) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.;

Redaschi, N.; Yeh, L. S. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **2004**, *32*, D115.

(19) Martens, L.; Vandekerckhove, J.; Gevaert, K. DBToolkit: processing protein databases for peptide-centric proteomics. *Bioinformatics* **2005**, *21*, 3584.

(20) Burkhart, J. M.; Premsler, T.; Sickmann, A. Quality control of nano-LC-MS systems using stable isotope-coded peptides. *Proteomics* **2011**, *11*, 1049.

(21) Storey, J. D.; Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **2003**, *100*, 9440.

(22) Everett, L. J.; Bierl, C.; Master, S. R. Unbiased statistical analysis for multi-stage proteomic search strategies. *J Proteome Res* **2010**, *9*, 700.

(23) Colaert, N.; Degroeve, S.; Helsens, K.; Martens, L. Analysis of the Resolution Limitations of Peptide Identification Algorithms. *J Proteome Res* **2011**.

(24) Fiala, G.; Stetter, K. O. *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. *Archives of Microbiology* **1986**, *145*, 56.

(25) Trauger, S. A.; Kalisak, E.; Kalisiak, J.; Morita, H.; Weinberg, M. V.; Menon, A. L.; Poole, F. L., 2nd; Adams, M. W.; Siuzdak, G. Correlating the transcriptome, proteome, and metabolome in the environmental adaptation of a hyperthermophile. *J Proteome Res* **2008**, *7*, 1027.

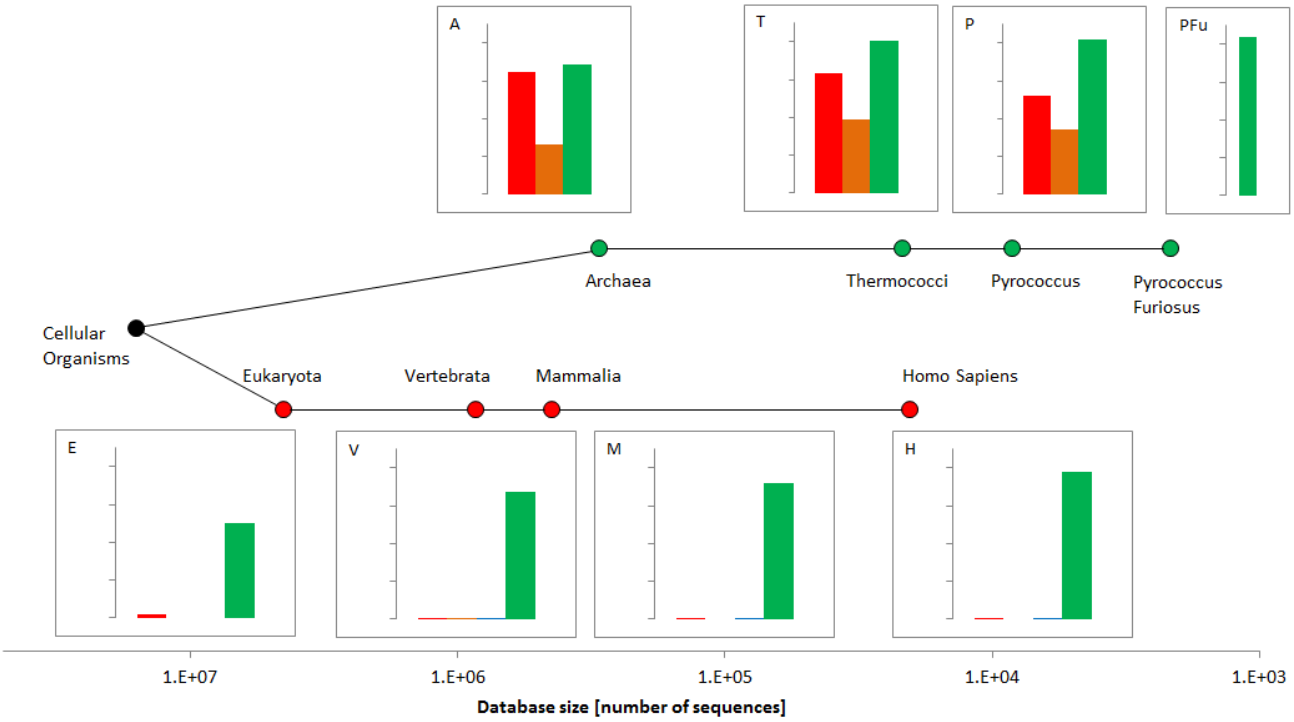
(26) Menon, A. L.; Poole, F. L., 2nd; Cvetkovic, A.; Trauger, S. A.; Kalisiak, E.; Scott, J. W.; Shanmukh, S.; Praissman, J.; Jenney, F. E., Jr.; Wikoff, W. R.; Apon, J. V.; Siuzdak, G.; Adams, M. W. Novel multiprotein complexes identified in the hyperthermophilic archaeon Pyrococcus furiosus by non-denaturing fractionation of the native proteome. *Mol Cell Proteomics* **2009**, *8*, 735.

(27) Cvetkovic, A.; Menon, A. L.; Thorgersen, M. P.; Scott, J. W.; Poole, F. L., 2nd; Jenney, F. E., Jr.; Lancaster, W. A.; Praissman, J. L.; Shanmukh, S.; Vaccaro, B. J.; Trauger, S. A.; Kalisiak, E.; Apon, J. V.; Siuzdak, G.; Yannone, S. M.; Tainer, J. A.; Adams, M. W. Microbial metalloproteomes are largely uncharacterized. *Nature* **2010**, *466*, 779.

(28) Im, Y. J.; Ji, M.; Lee, A.; Killens, R.; Grunden, A. M.; Boss, W. F. Expression of Pyrococcus furiosus superoxide reductase in Arabidopsis enhances heat tolerance. *Plant Physiol* **2009**, *151*, 893.

(29) Robb, F. T.; Maeder, D. L.; Brown, J. R.; DiRuggiero, J.; Stump, M. D.; Yeh, R. K.; Weiss, R. B.; Dunn, D. M. Genomic sequence of hyperthermophile, Pyrococcus furiosus: implications for physiology and enzymology. *Methods Enzymol* **2001**, *330*, 134.

SYNOPSIS TOC



This work presents the possibility of using the *Pyrococcus furiosus* proteome as an optimal standard to evaluate proteomic identification workflows. By exploring increasing evolutionary distance from the *Pyrococcus furiosus* to the *Homo sapiens* branch, we demonstrate how false positive identifications can be increasingly accurately discriminated while providing a sufficient quantity of proteins to ensure statistical significance.