Published in final edited form as:
J Proteome Res. 2012 December 7; 11(12): 5586–5591. doi:10.1021/pr300426s.

Recognizing uncertainty increases robustness and reproducibility of mass spectrometry-based protein inferences

Oliver Serang,
Department of Neurobiology, Harvard Medical School, Children's Hospital Boston, Boston, MA, USA, Oliver.Serang@Childrens.Harvard.edu

Luminita Moruz,
Stockholm University, Science for Life Laboratory, Department of Biochemistry and Biophysics, Solna, Sweden, luminita.moruz@scilifelab.se

Michael R. Hoopmann, and
Institute for Systems Biology, Seattle, Washington, Michael.Hoopmann@systemsbiology.org

Lukas Käll
Royal Institute of Technology(KTH), Science for Life Laboratory, School of Biotechnology, Solna, Sweden, lukas.kall@scilifelab.se

Abstract

Parsimony and protein grouping are widely employed to enforce economy in the number of identified proteins, with the goal of increasing the quality and reliability of protein identifications; however, in a counterintuitive manner, parsimony and protein grouping may actually decrease the reproducibility and interpretability of protein identifications. We present a simple illustration demonstrating ways in which parsimony and protein grouping may lower the reproducibility or interpretability of results. We then provide an example of a data set where a probabilistic method increases the reproducibility and interpretability of identifications made on replicate analyses of Human Du145 prostate cancer cell lines.

Introduction

Reproducible protein identifications, which are replicated in the various different biological fluids, are of paramount importance to the characterization of proteomes and for biomarker discovery; however, in shotgun proteomics the detected evidence is the fragmentation spectra, which come from peptides and not from proteins. To derive proteins, we are therefore dependent on protein inference procedures, which take protein databases as input. These databases hypothesize protein sequences that are anticipated to be either present or absent in the sample. The task is reminiscent of classical experimental design: we may test the hypothesized absence or presence of each individual protein, and reject the hypotheses based on how improbable the individual hypotheses are given the mass spectrometry data. The main advantage of such a probabilistic approach is that it can model the inherent uncertainty of the protein inference process. Unfortunately, few proteomics labs currently follow such a workflow for protein inference. Here, we present our main arguments why other labs would benefit from probabilistic protein inference.

In an effort to formalize the standards for presentation and publication of protein identifications the Paris guidelines[1] encouraged two alternative approaches intended to increase the reliability of protein identifications: protein grouping and parsimony. Today, parsimony and protein grouping are widely used in standard, well-regarded analysis

tools[2–4]; however, surprisingly, parsimony and protein grouping may actually lower the reproducibility and interpretability of protein identifications.

Parsimony

The advent of parsimony, which chooses the smallest set of proteins (or protein groups) accounting for a set of confidently identified peptides, comes from a well-intended attempt to quell the publication of massive numbers of identified proteins that were not reproducible in follow-up experiments[5]. By enforcing economy in the protein identifications (in statistics this is commonly referred to as "regularization"[6]), proteins supported by peptide evidence that could potentially be explained by other proteins are avoided. Today, parsimony is generally thought to yield conservative identifications, achieving an increased specificity at the cost of a lowered sensitivity; however, the certainty with which parsimony assigns presence and absence to proteins can actually lower the accuracy.

Here we present a simple example where parsimony lowers the reproducibility of the identified proteins. Figure 1 depicts three cases. In case 1, two peptides (peptides 2 and 3) are identified, all of which match three proteins (proteins A, B, and C). In case 2 all proteins have identical peptide evidence, except for an additional identified peptide (peptide 1), which matches only protein A. Likewise, in case 3 all proteins have identical peptide evidence, except for an additional identified peptide (peptide 4), which matches only protein C.

In case 1, any attempt to yield a single protein identification must result in a coin toss, choosing from the set {A, B, C} in an arbitrary fashion. In proprietary software packages the mechanism behind this choice is often opaque. Furthermore, two replicate experiments may result in case 2 (identifying protein A) and case 3 (identifying protein C). Despite the substantial overlap in peptide-level evidence, a parsimony-based approach will yield little reproducibility in the protein identifications from these replicates. Once again a coin toss is introduced, allowing a small bit of peptide evidence to completely tip the scales and change the result. The outcome is that parsimony lowers the reproducibility of protein identifications.

In general, "fixed threshold" methods (such as parsimony), use evidence to threshold peptides or proteins into strict present/absent states. Fixed-threshold methods employ hard-edge decision making, which considers only one conclusion from a set of possible outcomes. For example, the experimenter first derives a list of peptides (*e.g.*, using a target-decoy strategy with a fixed false discovery rate threshold), and then uses this list as an input to subsequent protein inferences. As noted by others[7], this initial threshold results in the experimenter filtering away peptide identifications that otherwise could be used for protein inference.

Grouping

These hidden complications from parsimony are further confounded by a tendency to incorrectly mix-up parsimony with "protein grouping". Parsimony resolves shared and ambiguous evidence by choosing the smallest set of proteins. In contrast, protein grouping merges several proteins into a single inference question concerning those proteins. Several grouping strategies are widely employed. For example, one grouping strategy would group only proteins that map to identical peptides, while another strategy would group together two proteins when the peptides from the first protein form a subset of the peptides from the second. In the paragraphs below we will present two reasons why protein grouping should be avoided, at least in the form in which it is practiced today. The first reason is that protein grouping presents a bad experimental design in the sense that it does not allow the

experimenter to compensate for known sources of variability; the second reason is that it hinders reproducible experiments.

Identification of a protein group answers a fundamentally different question than identification of a single protein: identifying protein A answers the question “Is protein A in the sample?” while identifying protein group {A, B, C} answers the question “Is at least one protein from the set {A, B, C} in the sample?” Grouping has traditionally been applied *after* the data (*i.e.* the peptide identifications) are observed. From a statistical viewpoint, it is considered good practice to formulate the tested hypothesis *before* data are observed, and any experimental design implying that questions are posed *after* the experimenter observes the data is discouraged. For this reason, unless multiple testing correction is applied, protein grouping should only be performed *before* spectra or other experimental evidence are considered. We will come back to how grouping prior to experiments can be achieved.

Grouping after the fact introduces a scenario where many multiple hypotheses are tested, and frequently yields replicate experiments with results that answer fundamentally different questions. For example, if proteins with identical peptide sets are grouped, then case 2 in Figure 1 would group {B, C} and leave protein A ungrouped, while case 3 would group proteins {A, B} and leave protein C ungrouped. There is no way to properly compare or merge the results because grouping was performed after the fact, and so the experiments answer different identification questions.

Probabilistic approaches

Fixed threshold methods like parsimony yield results that are inherently binary; they provide predicted event outcomes, but do not provide any information about the confidence or uncertainty in these predictions. There is no middle ground: a peptide is either identified or not identified, and proteins are subsequently identified or not identified. In contrast, probabilistic models[8;9] assign probabilities, which are interpretable measures of confidence, to wether a peptide or protein is present in the sample or not.

Because they do not threshold events into binary outcomes, probabilistic approaches can make use of *all* evidence, not simply the evidence that’s exceeded a certain threshold. As a result, probabilistic methods have access to greater amounts of information compared to fixed threshold methods. Furthermore, fixed threshold methods exhibit discontinuities in their response to minor changes to inputs (effectively, they employ a shifted heavyside function to transform the input data). As a result, a very minor change to a spectrum (for example, spectrum file format conversion resulting in a slight change in the precision of real values) can result in a disproportionately large change in the protein identifications. Probabilistic methods can smoothly respond to minor changes with correspondingly minor variations in the results.

Lastly, the confidence measures computed by probabilistic methods can always be thresholded *later* to derive a list of identified proteins. Thresholding as a last step preserves the extra information carried in probabilities. Furthermore, the probabilities can be used to assess the global and local false discovery rates (the latter is sometimes called the “posterior error probability”).

Caveats to probabilistic approaches

Due to the reasons described above, truly probabilistic methods (*i.e.* methods that compute the *actual* probabilities of peptides and proteins) are better tools than qualitative heuristics like parsimony and other fixed threshold methods. However, it should be noted that computing exact probabilities is a difficult task. Probabilistic methods are relying on a

number of assumptions to model the mass spectrometry process. When these assumptions are not realistic, their application is reminiscent of an incorrect hard-edge decision, and the resulting probabilities may become biased or skewed.

Some published attempts to create truly probabilistic methods implicitly make unrealistic assumptions, and are anecdotally known to exhibit a behavior similar to fixed threshold methods (*i.e.* they work very well in some settings but occasionally perform very poorly). One common hallmark of such behavior is a large number of peptides and proteins assigned probabilities of 0 and 1. The quintessential motivation of probabilistic approaches is that they respect the uncertainty inherent in inference; models that numerically favor certain results (*i.e.* probabilities of 0 and 1) do not comply with this, and thus are not substantially more useful than heuristics and fixed threshold methods. Likewise, numeric methods derived in an opaque manner are frequently rife with unrecognized assumptions. These methods behave probabilistically only when those assumptions are met, and so they are also not substantially superior to heuristics and fixed threshold methods. Recognizing and respecting uncertainty is of paramount import, even in the case of making assumptions.

For these reasons, we favor probabilistic methods that clearly state their assumptions. We also prefer simple methods over complex ones, whose inner workings are less transparent. These simple and clearly derived methods are frequently (but not necessarily) Bayesian in nature, because generative Bayesian models can describe the causal relationships in a mass spectrometry process in a simpler and more interpretable manner.

Methods

Sample preparation

Human Du145 prostate cancer cells were washed in cold PBS, and lysed in lysis buffer (8M urea, 0.1% RapiGest (Waters, USA), 100mM ammonium bicarbonate). Once lysed, the sample was diluted 8-fold with 100mM ammonium bicarbonate and protein concentration was measured by BCA assay. The proteins were denatured with 5 mM TCEP and free sulfhydryl bonds were alkylated with 10 mM iodoacetamide. The proteins were digested with trypsin. HCl was added to a final concentration of 50 mM and TFA was added to a final concentration of 1%. Peptides were desalted using Waters C18 Sep-Pak.

LC-MS/MS analysis

LC-MS/MS analysis was performed using a IntegraFrit (New Objective, USA) capillary (75 μm ID) packed with 20 cm of ReproSil Pur C18-AQ 3 μm beads (Dr. Maisch GmbH, Germany), and joined by union to a PicoTip (New Objective, USA) pulled silica tip (20 μm ID). Prior to loading the column, sample was loaded onto a fritted capillary trap (75 μm ID) packed with 2 cm of the same material. For each sample injection, 1 μg total protein was loaded onto the trap using an Agilent 1100 binary pump. Each sample was separated using a binary mobile phase gradient to elute the peptides. Mobile phase A consisted of 0.1% formic acid in water, and mobile phase B consisted of 0.1% formic acid in acetonitrile. The gradient program consisted of three steps at a flow rate of 0.3 μL/min using an Agilent 1100 nanopump: (1) a linear gradient from 5% to 40% mobile phase B over two hours, (2) a 10 minute column wash at 80% mobile phase B, and (3) column re-equilibration for 30 minutes at 5% mobile phase B.

Mass spectra were acquired on a LTQ Velos Orbitrap (Thermo Fisher Scientific) mass spectrometer operated on an 11-scan cycle consisting of a single high-resolution precursor scan event at 60,000 resolution (at 400 m/z) followed by ten data-dependent MS/MS scan events in the LTQ using collision induced dissociation (CID). The data-dependent settings were a repeat duration of 30 seconds, a repeat count of 2, and an exclusion duration of 3

minutes. Charge state rejection was enabled to fragment only 2+ and 3+ ions. The datasets are available in raw format via ProteomeCommons.org Tranche, using the following hashes:

```
ZqZiUrS98AyU1hYCu76AMNKtkl2x7soToeL8xMQtLoJnR+
Oh48DRF7CBtO6hv+84dCFxqmXICpd/0R1UZXM6/
IDApggAAAAAAAACWw==
CpQQAfZEh65p4EkUZBLZpCZ65K1PEw90EBt/j6pS/
mnf+shkDfuK78+tLtHSRfIkyB0OnUGuusi/+gUx9XSy+
p2pHxEAAAAAAAACXg==
```

Data processing

The MS1 data were deconvolved and the monoisotopic mass and charge state of the analytes were determined using Hardklör v2.01[10]. Next, we used Krönik v2.02 to determine those features that were observed in at least five consecutive scans, with a gap tolerance of one scan.

The fragmentation spectra were searched with Crux v1.35[11], using the `sequest-search` command. We used a precursor mass window of ±10ppm, and no missed cleavages were allowed. The datasets were searched against the 20185 protein sequences of the human Swiss-Prot 2011_09 database. All the datasets were also searched against a decoy database obtained by reversing the protein sequences from the target database. We used Bullseye[12] v1.3 to assure that the retention time of each peptide was assigned to the apex of its corresponding feature.

The resulting datasets were post-processed using Percolator v2.01[13]. Each replicate was processed separately using the two different inference protocols. Parsimony found the smallest set of proteins explaining the peptides identified at a 1% peptide-level q-value threshold. In the probabilistic method, protein posterior probabilities were calculated using Fido[14]. Protein q-values were calculated as the average posterior error probabilities of all proteins scoring as well or better than the current protein[15]. The probabilistic method used a 1% q-value protein-level threshold. Figure 2, compares the number and percent of proteins found by parsimony and a probabilistic approach in three, two, or one of the three replicates; the probabilistic approach identifies more proteins (both by number and percentage) reproduced in all three replicates.

Comparison of parsimony and probabilistic approach

To test our claim that parsimony lowers the reproducibility as compared to probabilistic methods we provide a simple comparison over three replicate analyses of Human Du145 prostate cancer cells. For each replicate we first searched our data with Crux[11] and Percolator[13], and subsequently inferred lists of proteins with both parsimony and a probabilistic method. We subsequently evaluated each of the two protein inference strategies by how concordant its inferred protein lists were over the replicates.

The first strategy was to deduce the most parsimonious set of proteins explaining the peptides at a 1% peptide-level q-value threshold. Although parsimony is equivalent to a, generally speaking, computationally prohibitive problem (the NP-hard set cover problem), it is in practice possible to solve for the graphs produced from a fairly specific digestion (trypsin, in this case).

The second strategy was to infer proteins by a Bayesian probabilistic method, Fido[14]. Fido generatively models the mass spectrometry process using a Bayesian network of noisy-OR

nodes. The Fido method marginalizes to compute protein and protein group-level probabilities by effectively summing over all possible sets of proteins present. Although we chose to use this particular probabilistic model, the general idea presented (*i.e.* recognizing uncertainty) is more general than this specific model.

Using the Fido method, we demonstrate that parsimony slightly lowers the reproducibility of the proteins identified (Figure 2) at a thresholds for the two methods that accept a similar number of identified proteins.

Discussion and recommendations

The predominant strategies regarding parsimony and protein grouping arose in an attempt to make protein identifications more reliable and to keep researchers from a more-identifications-is-better mentality; however, it is not the inherent uncertainty of protein inference that results in spurious or unreliable conclusions. Instead, it is only when this inherent uncertainty is met with unstable or overconfident assumptions that we decrease our ability to reliably characterize the contents of a protein sample. Viewed through this lens, parsimony and protein grouping actually arrive at overly concrete conclusions, making implicit assumptions such as, “When the peptides from one protein form a subset of the peptides from a second protein, the first protein is *never* present.” This rule of thumb may be generally true, but not universally true; the uncertainty should be reflected in the results by avoiding the all-or-nothing thresholding used by parsimony.

From a probabilistic point of view, the peptide evidence in case 1 in Figure 1 may strongly suggest that at least one of the proteins is present, it does not make a strong case that any one *particular* protein is present. This uncertainty, although not desirable, must simply be accepted as fact. Similarly, in case 2 protein A will have a higher probability, and B and C will have lower probabilities, but not enough to be absolutely certain that A is present and B and C are absent. Probabilistic methods are the only approaches that appropriately quantify this uncertainty. Thus, using truly probabilistic methods with assumptions that model the actual experimental setup is the best way to recognize and quantify the uncertainty inherent in protein identification. In the past 10 years, several probabilistic methods for protein identification have been published[8;9]. Although the current generation of probabilistic models are far from perfect, there exist several methods that are make relatively few assumptions where truth is unknown and are simple enough to be employed without great caution. As probabilistic models grow in popularity, more advanced and sophisticated methods will certainly follow. These methods will undoubtedly enable and advance mass spectrometry by bottling some skills from the experts who design them. In the past twenty years, many similar advances have been made in fields like image analysis or web search, and these advances have enabled a far greater reach of experts by allowing them to design models used by thousands of researchers.

However, as probabilistic models progress, we recommend that they frankly disclose their shortcomings. Mass spectrometrists have understandingly become increasingly resistant to believe the hype typically promulgated by new computational and statistical methods. By honestly stating and illustrating the shortcomings of statistical methods, we can earn the trust of those who use our tools. In review, we must reward methods that disclose and discuss their own flaws and weaknesses; these methods show greater respect for the uncertainty inherent to mass spectrometry-based inferences.

Likewise, unless multiple hypothesis testing is performed, protein grouping should be performed before the experiment. This can be performed by merging entries in the protein database that are similar. For example, proteins could be grouped if they stem from the same

or a similar gene. Alternatively, proteins could be grouped if they contain sufficiently overlapping peptide sets or if the proteins have a high degree of sequence similarity (as judged by a pairwise alignment). Grouping can be done in any way the user wants, as long as it is done before the experiment. At the very least, protein grouping must be performed consistently between replicate experiments in order to give a comparable and interpretable meaning to groups.

The Paris guidelines significantly advanced the field by first recognizing and addressing important pitfalls that arise when identifying proteins or reporting results; however, for the continued growth of mass spectrometry-based proteomics, we must continue to look critically on accepted practices and their unintended consequences and formalize the questions of genuine interest.

Hard-edge decision making can be beneficial when the conclusion drawn is almost certainly correct. In this case, probabilistic methods and hard-edge decisions will agree (probabilities will underscore the certainty of the outcome); however, in many aspects of fields like proteomics, this sort of certainty simply does not exist. As a field, we do not *truly* believe that two high-scoring peptide identifications certainly imply a present protein; likewise, we do not genuinely believe that a subset protein (*e.g.* protein B in case 2) is certainly not in the sample. Good decision making, whether probabilistic or absolute, occurs when we make assumptions only where we truly believe them, and where we are self-critical and forthcoming about the assumptions that we make from convenience, but do not fully endorse.

When, despite underlying uncertainty, hard-edge decisions must be made, we recommend making them at the last possible moment. As error (*e.g.* from an incorrect assumption or presumption of certainty) propagates, it compounds[16]. In the case of protein inference, a small amount of error from hard-edge decisions at the peptide level can result in much increased error at the protein level. When hardedge decisions cannot be avoided, it is preferable to apply them at the protein level, and thus prevent them from spreading and growing into pernicious errors.

Uncertainty itself is not the enemy; it is only when we make overreaching conclusions in the face of uncertainty that harm is done. A community that rewards the frank acknowledgment of the uncertainty in protein identification will itself be rewarded. Including probabilistic analyses in our practices and standards for publication will encourage us to be honest with ourselves about ambiguous results, and help distinguish these from highly informative, reproducible discoveries.

Acknowledgments

This work was supported with by a grant from the Swedish Research Council, federal funds from the National Institute of Health (grant No. T32NS007473) and the National Science Foundation MRI (grant No. 0923536), the Luxembourg Centre for Systems Biomedicine, and the University of Luxembourg. Additional thanks to Chung-Ying (Alan) Huang for supplying Du145 protein digests, and to Hanno Steen and Judith Steen for their comments.

References

1. Bradshaw, Ralph A.; Burlingame, Alma L.; Carr, Steven; Aebersold, Ruedi. Reporting protein identification data. Molecular & Cellular Proteomics. 2006 May; 5(5):787–788. [PubMed: 16670253]
2. Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. Journal of Proteome Research. 2007; 6(9):3549–3557. [PubMed: 17676885]

3. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Analytical Chemistry. 2003; 75:4646–4658. [PubMed: 14632076]

4. Li, YF.; Arnold, RJ.; Li, Y.; Radivojac, P.; Sheng, Q.; Tang, H. A Bayesian approach to protein inference problem in shotgun proteomics. In: Vingron, M.; Wong, L., editors. Proceedings of the Twelfth Annual International Conference on Computational Molecular Biology, volume 12 of Lecture Notes in Bioinformatics; Springer; Berlin, Germany. 2008. p. 167-180.

5. Carr, Steven; Aebersold, Ruedi; Baldwin, Michael; Burlingame, Al; Clauser, Karl; Nesvizhskii, Alexey. The need for guidelines in publication of peptide and protein identification data. Molecular & Cellular Proteomics. 2004; 3(6):531–533. [PubMed: 15075378]

6. Hastie T, Rosset S, Tibshirani R, Zhu J. The entire regularization path for the support vector machine. Journal of Machine Learning Research. 2004; 5:1391–1415.

7. Bern M, Kil YJ. Two-dimensional target decoy strategy for shotgun proteomics. Journal of Proteome Research. 2011; 10(12):5296–5301. [PubMed: 22010998]

8. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Journal of Proteomics. 2010; 73(11):2092–2123. [PubMed: 20816881]

9. Serang O, Noble WS. A review of statistical methods for protein identification using tandem mass spectrometry. Statistics and Its Interface. 2012; 5(1):3–20. [PubMed: 22833779]

10. Hoopmann MR, Finney G, MacCoss MJ. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics datasets using high-resolution mass spectrometry. Analytical Chemistry. 2007; 79:5620–5632. [PubMed: 17580982]

11. Park CY, Klammer AA, Käll L, MacCoss MP, Noble WS. Rapid and accurate peptide identification from tandem mass spectra. Journal of Proteome Research. 2008; 7(7):3022–3027. [PubMed: 18505281]

12. Hsieh E, Hoopmann M, Maclean B, Maccoss M. Comparison of database search strategies for high precursor mass accuracy MS/MS data. Journal of Proteome Research. 2009 Nov.

13. Käll L, Canterbury J, Weston J, Noble WS, MacCoss MJ. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. Nature Methods. 2007; 4:923–925. [PubMed: 17952086]

14. Serang O, MacCoss MJ, Noble WS. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. Journal of Proteome Research. 2010; 9(10):5346–5357. [PubMed: 20712337]

15. Granholm V, Käll L. Quality assessments of peptide–spectrum matches in shotgun proteomics. Proteomics. 2011; 11(6):1086–1093. [PubMed: 21365749]

16. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Molecular and Cellular Proteomics. 2009; 8(11):2405–2417. [PubMed: 19608599]

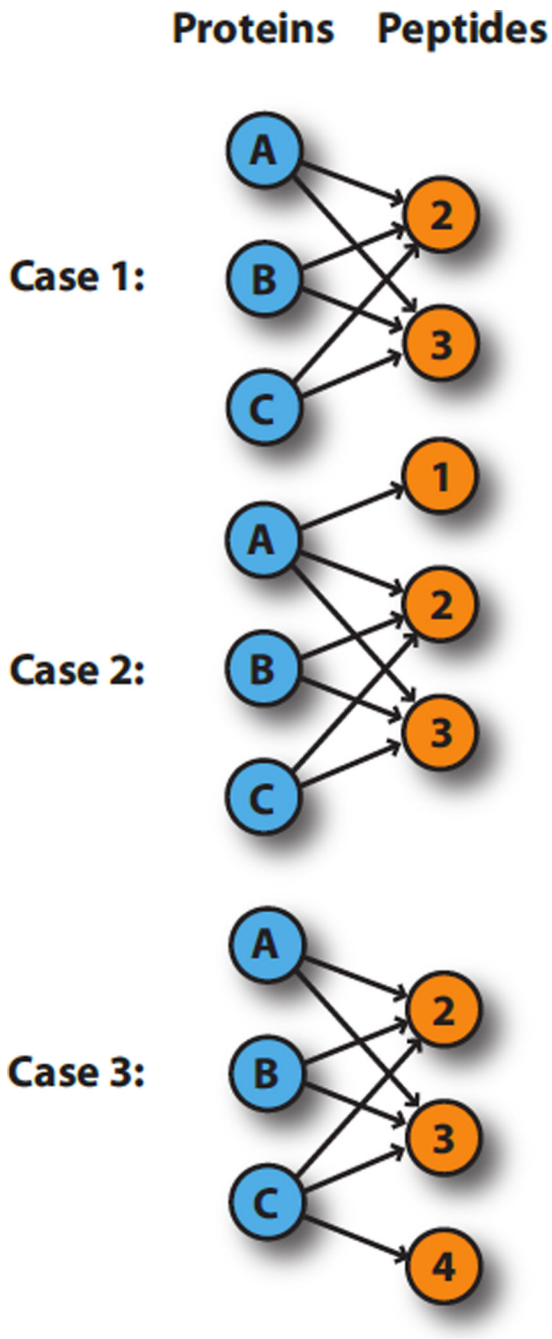

Figure 1. Three possible outcomes for a proteomics experiment

Each case is depicted as a Bayesian network (arrows represent causal dependencies in the experiment, by which proteins create peptides). **(Case 1)** Proteins A, B, and C are ambiguously identified by peptides 2 and 3. **(Case 2)** Identical to case 1, but with one additional identified peptide, mapping to protein A. **(Case 3)** Identical to case 1, but with one additional identified peptide, mapping to protein C.

	Identification method			
	Parsimony		Probabilistic	
Protein is found in	#	%	#	%
3/3 replicates	1332	64	1365	70
2/3 replicates	304	15	263	13
1/3 replicates	432	21	329	17

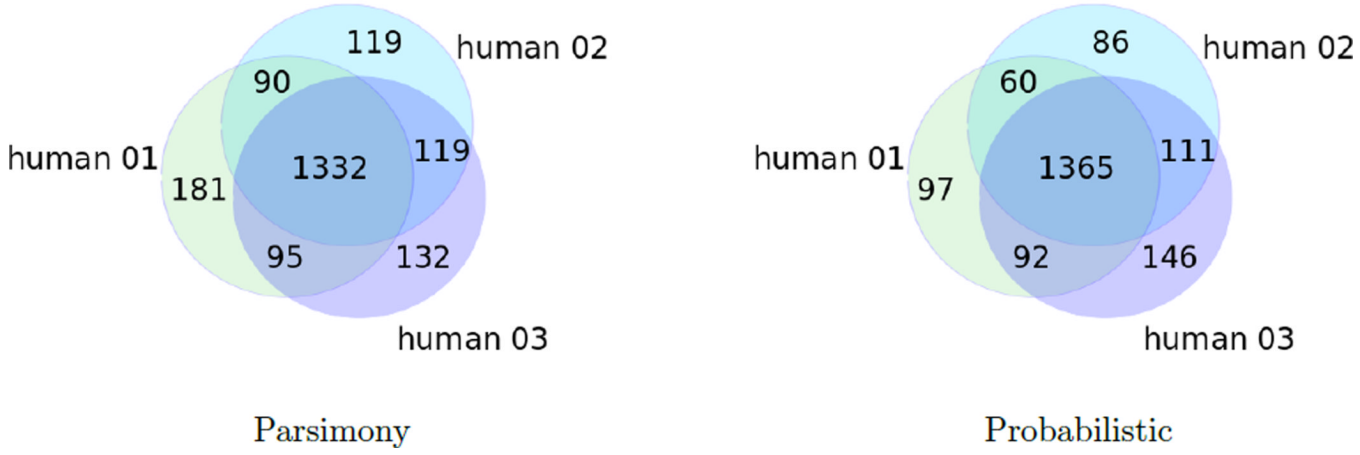

Figure 2. Overlapping protein identifications from three analyses of the same sample

For both methods we counted the number of proteins that were found in all (3/3 replicates), two (2/3 replicates), or exactly one (1/3 replicates) of the replicates. The relationship between the sets of proteins identified over the replicates are also shown in form of Venn diagrams when using parsimony and a probabilistic model. We find that the probabilistic method finds more proteins that are common to all replicates, and less identifications that are unique to one replicate, compared to parsimony.