

TECHNICAL BRIEF

The minotaur proteome: Avoiding cross-species identifications deriving from bovine serum in cell culture models

Jakob Bunkenborg^{1*}, Guadalupe Espadas García², Marcia Ivonne Peña Paz³, Jens S. Andersen¹ and Henrik Molina²

¹ Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark

² Centro de Regulación Genómica (CRG), Barcelona, Spain

³ Universitat Pompeu Fabra, Barcelona, Spain

Cell culture is a fundamental tool in proteomics where mammalian cells are cultured *in vitro* using a growth medium often supplemented with 5–15% FBS. Contamination by bovine proteins is difficult to avoid because of adherence to the plastic vessel and the cultured cells. We have generated peptides from bovine serum using four sample preparation methods and analyzed the peptides by high mass accuracy LC-MS/MS. Distinguishing between bovine and human peptides is difficult because of a considerable overlap of identical tryptic peptide sequences. Pitfalls in interpretation, different database search strategies to minimize erroneous identifications and an augmented contaminant database are presented.

Received: February 16, 2010

Revised: May 25, 2010

Accepted: June 1, 2010

**Keywords:**

Artifacts / Bioinformatics / Cell culture / MS / SILAC

Cell culture models have been pivotal in fundamental studies on the cellular regulation and development and have provided many insights into the inner workings of human cells. Protein biomarkers for early diagnosis of disease have very often escaped detection because proteins of diagnostic value often originate from relatively small heterogeneous cell lesions that are diluted into large volumes of highly complex body fluids present in a large dynamic range of protein abundances. To circumvent these problems, numerous studies depart from cell monoculture models to identify alterations in cell membrane proteins or secreted proteins that can serve as serological markers. Traditionally, mammalian cells are cultured *in vitro* using a chemically defined medium (such as Eagle's modified essential medium) that is supplemented with between 5 and 15% FCS or FBS. The fetal serum supplies a rich protein solution that has low antibody content while containing growth factors, transport proteins and attachment factors that promote cell growth. A 10% FBS formulation adds around

5–6 mg protein *per* millilitre medium and although cells undergo extensive washing prior to harvest or collection of secreted proteins, it is difficult to completely remove the bovine proteins below the detection limits of modern MS equipment. Here, we evaluate database search strategies to reduce the number of false-positive identification of human proteins from a background of bovine serum proteins. We prepared peptides from bovine serum using different sample preparation methods, analyzed the samples by LC-MS/MS and probed the resulting data using commonly applied database search strategies. The Minotaur is a hybrid creature from Greek mythology with a bull's head on the body of a man. We demonstrate here that without careful experimental setup and data analysis, there is a likelihood of making mythological protein identifications from this creature.

To estimate the problem of cross species overlapping peptides, we derived *in silico* bovine and human tryptic peptides that could fall within our typical window of LC-MS/MS analysis. We commonly apply the data acquisition

Correspondence: Dr. Henrik Molina, Centro de Regulación Genómica (CRG), C/Dr. Aiguader 88, 08003 Barcelona, Spain

E-mail: henrik.molina@gmail.com

Fax: +34-93-316-00-99

*Additional corresponding author: Jakob Bunkenborg

E-mail: bunkenborg@bmb.sdu.dk

scheme that only multiple charged ions in the m/z range of 300–1500 Th are selected for fragmentation. Since we seldom identify peptides with more than 50 amino acid residues, we calculated the distribution for tryptic peptides with no missed cleavages with a mass greater than 600 Da and less than 50 amino acid residues long. For this analysis, we downloaded the human and bovine protein database (ftp://ftp.ebi.ac.uk/pub/databases/IPI/bovine v3.52 containing 31 522 entries and Human v3.67 containing 87 040 entries). The comparison of the human and bovine tryptic peptide distribution is shown in Fig. 1. As illustrated in Fig. 1B, the overlap based on peptide sequence identity is considerable even for long peptides and more than 20% of

the human peptides up to 25 amino acid residues has an identical “tryptic sequence” in the bovine database. A common strategy for avoiding erroneous identifications is to augment the database with a collection of common contaminants [1, 2] and since we use MaxQuant for the further data analysis, we selected the contaminant file (associated with version 1.0.13.13, 262 entries) and appended it to the databases using the SequenceReverser tool included in the MaxQuant package so that reversed sequence identifiers are tagged with REV_ and contaminant sequences identifiers are tagged with CON_.

To generate a large set of representative tryptic peptides from FBS, we used four different sample preparation

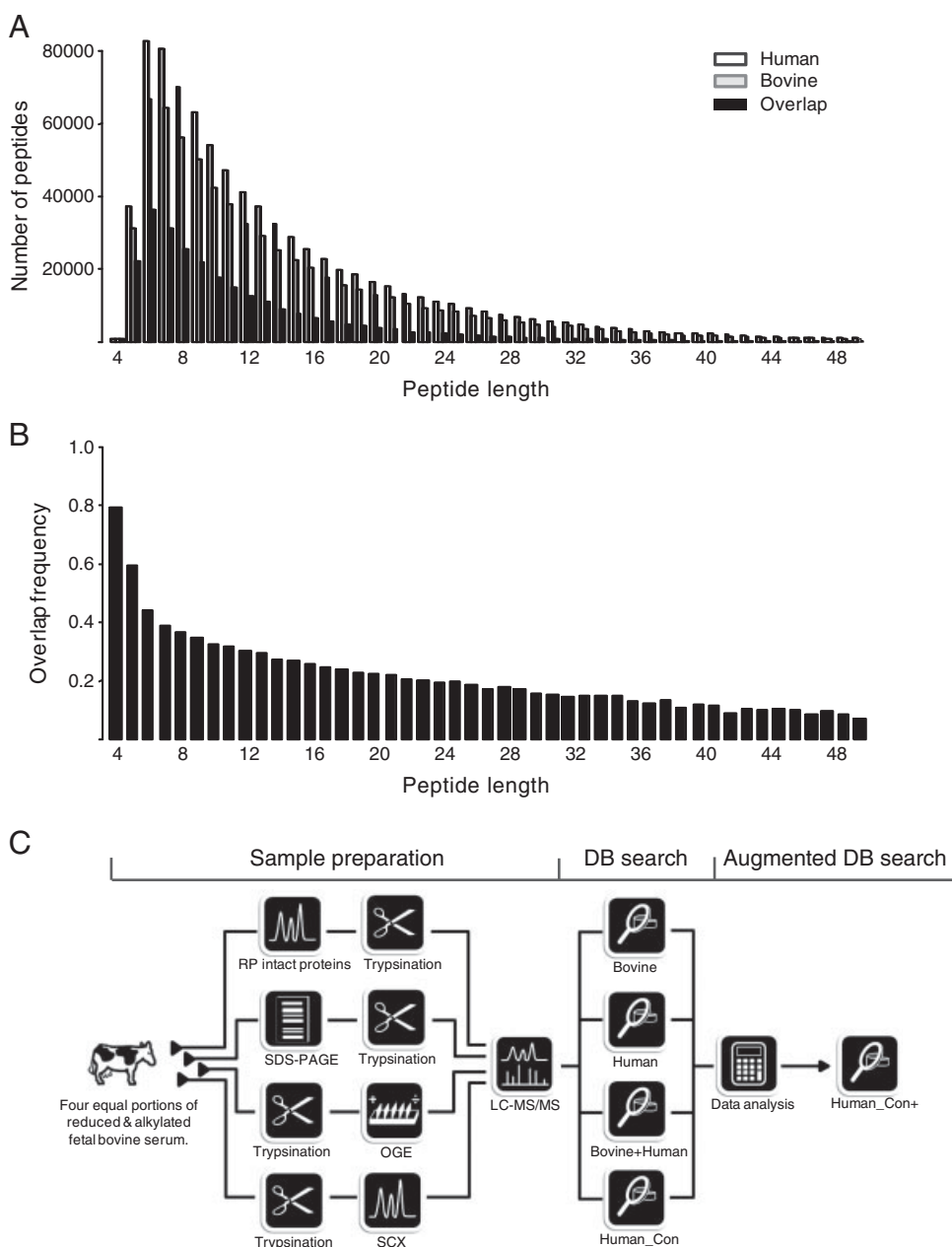


Figure 1. (A) Distribution of unique tryptic peptide sequences (no missed cleavage) derived from the human and bovine IPI databases with a mass above 600 Da. The subset of overlapping identical sequences is shown in black. (B) The relative frequency of identical sequences between the bovine and human database. Although the overlap between human and bovine peptides decreases as peptide length increases, there is still a considerable overlap. (C) Schematic representation of the workflow. To generate a sample that encompasses bovine contaminating peptides that can appear in human cell culture models, proteins from FBS were prepared in four different ways: Protein centric based on separation of intact proteins using either reversed phase chromatography (109 proteins identified) or SDS-PAGE (312 proteins identified); and peptide centric by proteolysis followed by either off-gel electrophoresis (292 proteins identified) or strong cation exchange (252 proteins identified). Cysteines were reduced and alkylated prior to separation; all proteins were digested with trypsin and the resulting peptides analyzed by high mass accuracy LC-MS/MS. MS/MS data were queried against human and bovine IPI databases.

techniques as summarized in Fig. 1C. A stock solution of reduced and alkylated fetal serum was generated by denaturing 200 μ L fetal serum (Gibco/Invitrogen) diluted five times in 25 mM ammonium bicarbonate (Fluka Analytical, Germany) and 1000 μ L 2,2,2-trifluoroethanol (Sigma-Aldrich, St. Louis, MO, USA) were added according to a protocol provided by Dr. Christine Miller of Agilent Technologies [3]. The denatured serum was reduced by adding 30 μ L 1 M DTT (Sigma-Aldrich), allowed to incubate 1 h at 60°C and hereafter alkylated with 70 μ L 1 M iodoacetamide (Sigma-Aldrich). Four equal portions of approximately 400 μ g FBS were separated into 24 fractions using either (i) macro-porous C18 reversed-phase separation of intact proteins (Macroporous RP-C18, Agilent, Santa Clara, CA, USA) with a 43 min gradient of 97% of 0.10% TFA/3% of 0.08% TFA in ACN to 40% of 0.10% TFA/60% of 0.08% TFA in ACN followed by trypsin digestion of the collected fractions [4]; or (ii) SDS-PAGE separation (NuPAGE[®] Novex 4–12% Bis-Tris Gel, Invitrogen) of the proteins followed by in-gel digestion; or (iii) in-solution digestion of the serum proteins followed by off-gel isoelectrofocusing on an Agilent 3100 according to the protocol of Hubner *et al.* [5]; or (iv) in-solution digestion of the serum proteins and fractionation using strong cation exchange chromatography (Poly-SULFOETHYL Aspartamide, PolyLC, Colombia, MD, USA) with a gradient of 97% of 10 mM KH₂PO₄, 25% ACN/3% of 10 mM KH₂PO₄, 300 mM KCl, 25% ACN increasing to 1% 10 mM KH₂PO₄, 25% ACN/99% of 10 mM KH₂PO₄, 300 mM KCl, 25% ACN in 50 min. Trypsin (Promega, Madison, WI, USA) was used for all protein digestion. Each

of the 24 fractions from the four strategies were analysed by high mass accuracy (Orbitrap XL, Thermo Fisher Scientific, Bremen, Germany) LC-MS/MS using a flow of 300 nL/min and a gradient increasing from 10% of 0.1% formic acid/90% of 0.1% formic acid, ACN to 55% of 0.1% formic acid/45% of 0.1% formic acid, ACN (Agilent 1200). The data acquisition cycle was a full scan (100 000 resolution) in the orbitrap (AGC target 5e5) followed by CID of the ten most intense multiply charged ions in the LTQ (MSn AGC target 1e4) with dynamic exclusion of the selected ions for 60 s. A single ion was used as lock mass. MS/MS spectra were processed with the Quant module of MaxQuant and the centroided data were queried against bovine and human IPI-decoy databases using MASCOT (version 2.2.05, Matrix-Science, London, UK). A total of 833 112 MS/MS spectra were queried. For the search queries, we used tryptic specificity and allowed three missed cleavages. The mass tolerance was 7 ppm for precursor ions and 0.5 Da for product ions. Carbamidomethylation of cysteine as a fixed modification and oxidation of methionine and N-terminal glutamate conversion to pyro-glutamate as variable modifications. The results were imported into MaxQuant using the Identify module allowing an FDR of 0.01 for both peptides and proteins.

The same set of MS/MS spectra was searched against five different databases: Bovine IPI, Human IPI, the combined Bovine and Human IPI, Human IPI with the MaxQuant contaminant list and finally the Human IPI with an augmented contaminant list that includes all the identified bovine proteins in this study. The protein and peptide

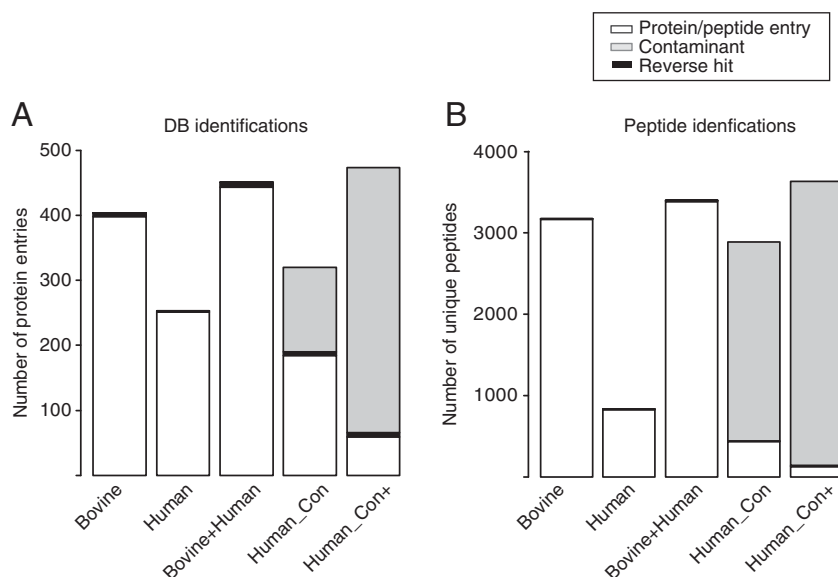


Figure 2. The minotaur proteome. (A) Protein identifications based on tryptic peptides derived from FBS when searched against the five different databases: bovine IPI, human IPI, a concatenated bovine and human IPI, human IPI augmented with a contaminant fasta-database and human IPI with a contaminant fasta-database enlarged with the bovine identifications. (B) Peptide identifications. When the bovine data was searched against the human database, 827 unique peptides were identified using an FDR of 1%. Incorporating contaminant entries into the human IPI database increased the number of identified peptides and decreased the number of bovine peptides identified as human.

identifications returned *via* MASCOT and MaxQuant are summarized in Fig. 2 (and the details of each identification can be found in Supporting Information Table 1). The majority of proteins identified in both the bovine and human databases are found among the most abundant serum proteins [6]. As anticipated by the *in silico* comparison of human and bovine tryptic peptides (Fig. 1A), a large number of spectra were matched to an identical peptide sequence when searching the non-augmented human and bovine databases. Sixty-nine percent of the spectra matching human peptides were also retrieved in the bovine db with the majority of these (85%) retrieving identical sequences suggesting that cross species sequence identity is a greater problem than sequence similarity.

The human protein identifications from bovine tandem mass spectra tend to be based only on one (46% of the proteins) or a few peptides, which might alarm the researchers that the human identifications could be dubious. However, the researcher might still accept such identifications since some of the “identified” human peptide sequences cannot be found in the bovine database leading to the false conclusion that the identifications are supported by uniquely human peptides. An example of this problem is illustrated in Fig. 3 where the same MS/MS retrieves two unique but very similar peptide sequences both from Complement C3. Since the identified human peptide is uniquely human and cannot be found in the bovine database, a simple peptide search in the bovine database would not reveal the erroneous assignment. One solution to account for bovine contaminants could be to expand the database to include both the human and bovine proteome but this is not

an ideal solution: it increases database size considerably, separating groups of identified IPI entries by species takes some scripting and most importantly perfectly valid and unique human peptides having an identical bovine peptide sequence could be discarded, although the peptide never has been observed in bovine serum. An alternative strategy is extending the database to include known contaminants using the contaminant list of MaxQuant. Common contaminants from multiple species can be included (*e.g.* commonly used proteases as porcine trypsin or lys-C from *Achromobacter lyticus*) by appending a small contaminant database to the database of interest. This increases the total number of identifications and reduces the number of false human identifications and conveniently tags the identification as a possible contaminant (Fig. 2, Human_Con). Augmenting the MaxQuant contaminant list with the additional bovine protein sequences identified by the four sample preparation strategies, we could further decrease human protein identifications (Fig. 2, Human_Con+). The remaining 59 human proteins identified include some most likely genuine human skin contaminants added during sample processing that are not yet part of the contaminant list: *e.g.* hornerin, dermcidin and galectin-7 that have been observed in a large number of 2-DE experiments of unrelated samples [7]. The fasta file used to supplement the MaxQuant contaminant list is supplied as the Supporting Information file: Contaminant_with_cow.fasta. By using this augmented database, human peptides with identical bovine sequences will only be tagged as contaminants, if they have been observed as a real component of bovine serum, so that identical sequences are innocent till proven guilty.

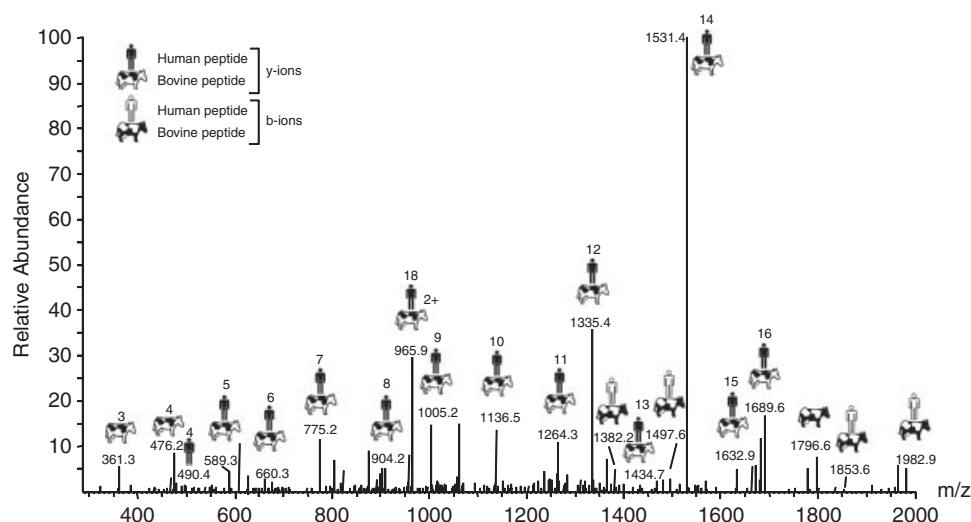


Figure 3. Sequence identity is not enough to eliminate human peptide identifications deriving from bovine peptides. As an example, the shown spectrum was identified as deriving from the human peptide ILLQGTTPVAQMTEDAVIDAER with a MASCOT score of 103 and illustrates how the MS/MS spectrum of the bovine peptide ILLQGTTPVAQMTEDAIDGER (MASCOT score 110) can lead to erroneous assignments. In the figure, γ - and β -fragment ions assigned to the bovine peptide and the human peptide are marked with intuitive symbols. In addition, γ -ions are numbered to appreciate the long series of consecutive fragments that fits both matches. The human peptide match with a convincing continuous γ -ion series would most likely pass through data evaluation as a rock-solid identification with a peptide sequence that cannot be found in the bovine database.

The difficulty in removing proteins from cultured cells by washing is partly due to adsorption of the proteins both to the cell and the polystyrene surfaces. A study on albumin adsorption to polystyrene [8] illustrates the magnitude of the problem. The surface was incubated with either 50 or 100 µg/mL serum albumin for 1 h, rinsed twice in Milli-Q water, placed with Milli-Q water for 4 h on a shaker and then rinsed twice again in Milli-Q water. The amount of albumin adsorbed to the surface after this extensive wash procedure was estimated to be 478 and 735 ng/cm² for the two albumin concentrations. This would suggest that it is not negligible amounts of protein that can be adsorbed to the >135 cm² surface of a 15 cm dish when incubated with the ~6 mg/mL protein of 10% FBS growth medium. Another confounding factor is that the polystyrene tissue culture plates from different vendors often are manufactured differently to enhance cell adhesion. A study [9] compared IL-1ra production by human cells in tissue plates from five different vendors and found that the brand with the lowest level of albumin adsorption also gave the lowest yield of IL-1ra. Analysing the proteins adhering to the plastic (*e.g.* by incubating empty plates with growth medium) will only partly reflect the contamination problem because the bovine proteins that adhere to the cell surfaces will result in a different landscape of “carry-over” proteins.

One strategy to eliminate bovine contaminants would be to remove all animal-derived components in the experimental workflow. However, moving to serum-free cell culture does not remove all contaminating factors because several proteins are required to promote growth. For example, insulin and transferrin are common supplements in serum-free cell culture, although these could be made recombinantly with very high purity [10]. Other sources of exogenous proteins could be the trypsin used for passaging the cells or collagens added to improve cell attachment. In this study, we show the benefits of using a list of contaminants experimentally derived from bovine serum. A complementary way of differentiating between proteins deriving from the cells and contaminating proteins is to use SILAC [11], since only proteins synthesized in the presence of stable isotope labelled amino acids can incorporate the labelled amino acids. If a comparison of several conditions is needed, heavier stable isotopes could be incorporated into all conditions (*e.g.* Arg6/Lys4 versus Arg10/Lys8) of if “natural isotopes” are required one could use label swapping to ensure that ratio differences do not derive from contaminants.

In summary, we have shown how bovine proteins can result in erroneous human protein identifications and suggest that this problem can be alleviated by augmenting databases with an extended list of experimentally obtained possible bovine serum protein contaminants.

J. B. acknowledges funding from the European Commission's 7th Framework Programme (grant agreement HEALTH-F4-2007-200767/APOSYS). M. I. P. P. is partly funded by ProteoRed (the Spanish Proteomics Network). The study is carried out in a laboratory affiliated with ProteoRed.

The authors have declared no conflict of interest.

References

- [1] Schandorff, S., Olsen, J. V., Bunkenborg, J., Blagoev, B. *et al.*, A mass spectrometry-friendly database for cSNP identification. *Nat. Methods* 2007, 4, 465–466.
- [2] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 2008, 26, 1367–1372.
- [3] Horth, P., Miller, C. A., Preckel, T., Wenz, C., Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. *Mol. Cell. Proteomics* 2006, 5, 1968–1974.
- [4] Molina, H., Horn, D. M., Tang, N., Mathivanan, S., Pandey, A., Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA* 2007, 104, 2199–2204.
- [5] Hubner, N. C., Ren, S., Mann, M., Peptide separation with immobilized pl strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics* 2008, 8, 4862–4872.
- [6] Anderson, N. L., Anderson, N. G., The human plasma proteome. *Mol. Cell. Proteomics* 2002, 1, 845–864.
- [7] Dumont, D., Noben, J. P., Raus, J., Stinissen, P., Robben, J., Proteomic analysis of cerebrospinal fluid from multiple sclerosis patients. *Proteomics* 2004, 4, 2117–2124.
- [8] Browne, M. M., Lubarsky, G. V., Davidson, M. R., Bradley, R. H., Protein adsorption onto polystyrene surfaces studied by XPS and AFM. *Surf. Sci.* 2004, 553, 155–167.
- [9] Clinchy, B., Youssefi, M. R., Hakansson, L., Differences in adsorption of serum proteins and production of IL-1ra by human monocytes incubated in different tissue culture microtiter plates. *J. Immunol. Methods* 2003, 282, 53–61.
- [10] Keenan, J., Pearson, D., Clynes, M., The role of recombinant proteins in the development of serum-free media. *Cytotechnology* 2006, 50, 49–56.
- [11] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B. *et al.*, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* 2002, 1, 376–386.