A Bayesian two-layer latent variable model for single-isoform proteogenomics inference

Simone Tiberi

Department of Statistical Sciences, University of Bologna, Bologna, Italy

1 Objectives, originality and innovation

Background. Studying protein diversity and identifying physiologically relevant protein isoforms are crucial steps in biomedical research. Biological mechanisms, such as alternative splicing or alternative promoter usage, allow single genes to code for multiple isoforms; for example, in humans, it was estimated that approximately 20,000 genes may give rise to over 300,000 isoforms. Nonetheless, **most experimental approaches detect proteins at the gene level**, and do not allow investigating proteome complexity at the isoform level.

Current state of research in the field. At present, the main strategy to infer proteins is via "bottom-up" proteomics, where proteins are indirectly measured via peptides, which act as surrogate markers for their protein(s) of origin. However, most peptides, called shared peptides, map to multiple protein isoforms; this usually results in ambiguous protein identifications (Figure 1), and **inferential results are typically abstracted to the gene level**. Although a few methods have been proposed to perform inference at the isoform level (notably, EPIFANY, FIDO, and PIA), due to the prevalence of shared peptides, protein detection is affected by low statistical power. Furthermore, inference only focuses on identifying proteins (presence vs. absence), and not on further measures such as protein abundance. Similarly, differential tools (e.g., DEP, MAP and DEqMS) comparing protein expression between experimental conditions (e.g., healthy *vs.* diseased) also provide results at the gene level.

Objectives. We propose a novel **statistical method for proteogenomics inference**, via integration of transcriptomics data, which is a prerequisite and correlate of protein abundance. We will **jointly model proteomics and transcriptomics data** in a Bayesian probabilistic framework, in order **to perform inference on individual protein isoforms**. Note that, our method may also be used with proteomics data alone, but accuracy is expected to increase when transcriptomics data is also provided. The overarching goal of this proposal is to develop a well-documented, and widely-used **open source software tool**, based on rigorous statistical methods and efficient computational strategies, distributed as a **Bioconductor R package**, which makes it easy to install and integrate with existing pipelines, and accompanied by an example vignettes that facilitates its usage.

Multiomics integration. First, transcriptomics data (e.g., RNA-sequencing) will be employed to estimate the relative abundance of transcript isoforms: these estimates will be used to formulate an **informative prior** for the relative abundance of the respective protein isoforms.

Two-layer latent variable model. Inference on protein isoforms is complicated by several technical limitations; in particular: i) peptides may be erroneously detected, even when absent; ii) shared peptides are compatible with multiple protein isoforms. We explicitly model these two sources of uncertainty, with a two-level latent variable model. First, we sample the presence/absence of each peptide based on its estimated probability of being mistakenly detected (provided directly by proteomics tools). Second, for shared peptides that were estimated as being present, we allocate their abundance across the protein isoforms they map to, hence recovering the presence and abundance of each protein isoform.

Future directions. Initially, our method will infer parameters from single samples. In a second stage, we will extend our model to jointly fit multiple samples (i.e., biological replicates) in a **Bayesian hierarchical framework**, hence allowing for sample-specific parameters, while sharing information across samples. This

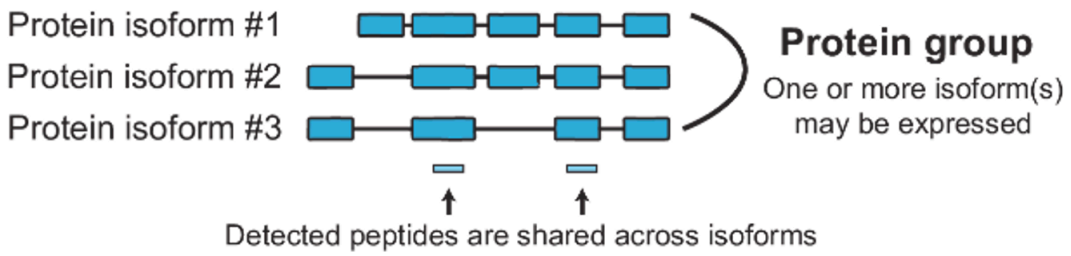

Figure 1: Example of two shared peptides (highlighted by the black arrows) across three protein isoforms.

extended model will also enable **differential testing between conditions** (e.g., treated *vs.* untreated), hence identifying the individual protein isoforms that vary across conditions. Furthermore, although current work involves analysis of bulk data (i.e., mean signal across multiple cells), our future methods will progress into **single-cell variants**. This would allow us to perform cell-type specific inference, hence investigating how protein isoforms change between cell types. Overall, we believe that the research direction presented here will lead to **several research products**.

Novelty. Our approach will not only **infer the presence/absence of protein isoforms**, but also **estimate their abundance**. Additionally, we will provide **a measure of the uncertainty of our estimates**, via the posterior probability the a protein isoform is present in the sample, and a posterior credible interval of its abundance. Importantly, to be best of our knowledge, at present, no tool infers protein abundance, jointly models multiple samples, and performs differential testing at the isoform level; furthermore, no method in proteomics has ever used a two-layer latent variable approach.

2 Impact, feasibility and implementation

Protein abundance. Given P protein isoforms, we assume that the overall protein abundance of a sample, denoted by $n \in \mathbb{N}$, is distributed across the P isoforms according to a multinomial distribution:

$$X = (X_1, \dots, X_P) \sim MN\left(\pi = (\pi_1, \dots, \pi_P),\, n = \sum_{p=1}^{P} x_p\right), \tag{1}$$

where X_p represents the random variable indicating the overall abundance originating from the p-th protein (and x_p is its realization), and π_p is the probability that a unit of abundance comes from the p-th protein, with $\sum_{p=1}^{P} \pi_p = 1$. The probability that the p-th protein is present is estimated via $Pr\left(X_p > 0\right)$. Nonetheless, although we aim to infer proteins, **measurements refer to peptides**; therefore, protein isoform abundances, in X, are treated as latent variables and sampled via a **data augmentation** approach.

Latent variables - 1^{st} layer. Assume that N peptides are detected in total, and that PE_i is the estimated probability (taken as input) that the i-th peptide, albeit absent, is erroneously detected.

We sample if a peptide has been wrongly detected following a Bernoulli distributions:

$$\epsilon_i | PE_i \sim Bern(PE_i), \text{ for } i = 1, \dots, N, \tag{2}$$

where $\epsilon_i = 1$ if the $i - th$ peptide has been mistakenly detected, and 0 otherwise.

Latent variables - 2^{nd} layer. We define Y_i as the abundance of the i-th peptide, and ψ_i as the list of proteins the i-th peptide maps to; we further denote by X_{pi} the (unknown) abundance of peptide i that comes from protein p. We assume that the abundance of the i-th peptide can be redistributed to the proteins in ψ_i according to a multinomial distribution:

$$(X_{1i}, \dots, X_{Pi}) | \pi, \psi_i, Y_i, \epsilon_i \sim MN\left(\tilde{\pi}^{(i)}, Y_i(1 - \epsilon_i)\right), \text{ for } i = 1, \dots, N, \tag{3}$$

where $\tilde{\pi}^{(i)} = \left(\tilde{\pi}_1^{(i)}, \ldots, \tilde{\pi}_P^{(i)}\right)$, with

$$\tilde{\pi}_p^{(i)} = \frac{\pi_p \mathbb{1}(p \in \psi_i)}{\sum_{p'=1}^{P} \pi_{p'} \mathbb{1}(p' \in \psi_i)}, \tag{4}$$

where $\mathbb{1}(A)$ is 1 is A is true, and 0 if A is false. In other words, $\tilde{\pi}_p^{(i)}$ is proportional to π_p if the i-th peptide maps to the p-th protein, and 0 otherwise. Note that, in (3), the peptide abundance, $Y_i(1-\epsilon_i)$, is 0 if the peptide has been sampled as mistakenly detected in (2) (i.e., when $\epsilon_i = 1$). The protein isoform abundances are then recovered by adding the abundances obtained from the N peptides allocations: $x_p = x_{p1} + \ldots x_{pN}$, for $p = 1, \ldots, P$.

Prior formulations. We use an informative Dirichlet prior for π:

$$\pi \sim Dir\left(\delta = (\delta_1, \ldots, \delta_P)\right), \tag{5}$$

with δ proportional to the transcript isoforms abundance. This is a **conjugate prior**, which results in a convenient posterior distribution:

$$\pi | x, \delta \sim Dir((x_1 + \delta_1, \ldots, x_P + \delta_P)). \tag{6}$$

We further assume a weakly-informative discrete uniform prior for the overall abundance of proteins:

$$X_p \sim Unif(0, 1, \ldots, n), \text{ for } p = 1, \ldots, P. \tag{7}$$

MCMC. Parameters and latent states are alternately sampled from their conditional distributions, following two **Gibbs samplers**, via a Metropolis-within-Gibbs Markov chain Monte Carlo (MCMC) scheme. Convergence is assessed via Heidelberger and Welchs stationarity test.

Benchmarking. We will design various benchmarks, on both real and simulated data, and evaluate the performance of our approach and several competitors. For our model evaluation, we will be **supported by our collaborators** at Dr. Sheynkman's lab (University of Virginia), who have already collected data, and will generate more as the project evolves.

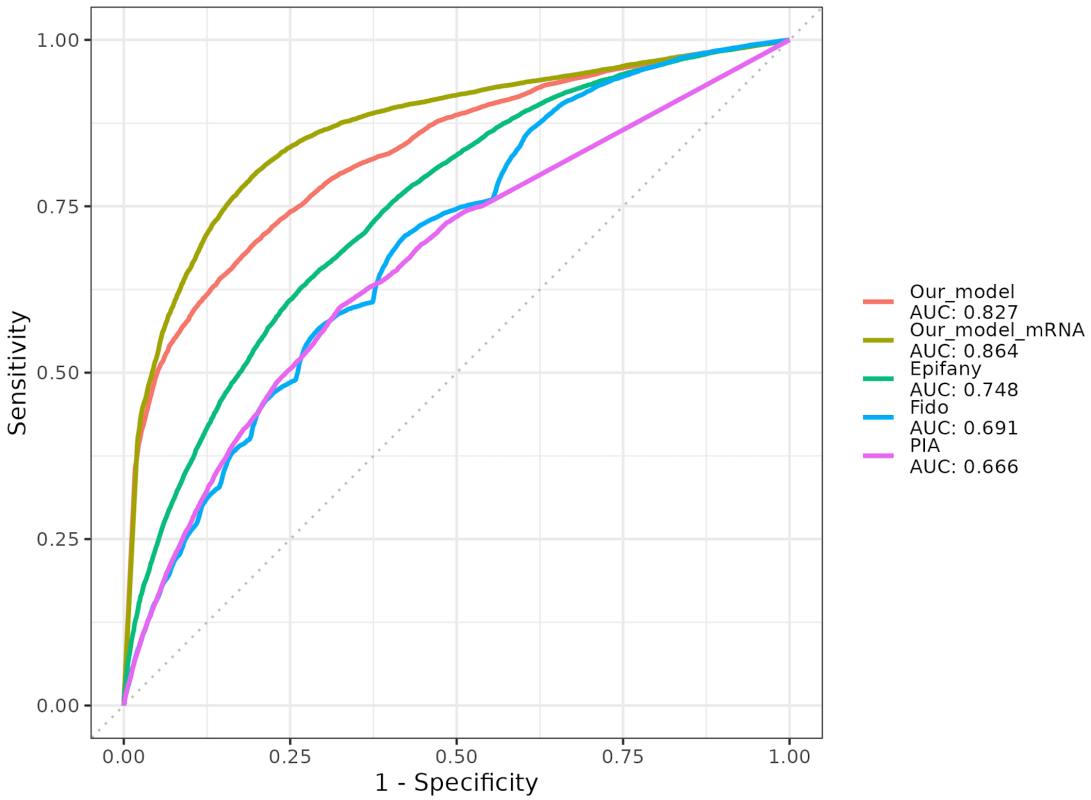

Figure 2: ROC curve for the detection of protein isoforms.

Preliminary results. We have implemented a prototype of our model and tested it on real data. In particular, our collaborators collected proteomics measurements from six distinct proteases: with a leave-one-out approach, we analyzed one at a time, and used the (unique peptides from the) remaining five to validate results. When using proteomics data alone, our approach ("Our_model"; AUC of 0.83) displays **higher sensitivity and specificity** than competitors (AUCs between 0.67 and 0.75) at detecting protein isoforms; this gap further increases when adding transcriptomics data ("Our_model_mRNA"; AUC of 0.86; Figure 2). In addition, our estimated **abundances highly correlate** (Pearson correlation of 0.72) with the corresponding ground truth (Figure 3). Finally, albeit full MCMC schemes can be computationally cumbersome, our algorithm was **efficiently coded in C++**, and ran in **∼2 minutes**.

Impact. This proposal aims at creating an **all-rounded statistical method for protein isoform inference**. Our tool could be of great utility to computational biologists, by unlocking a great unexploited potential for biological discoveries. For instance, it was shown that proteins of the transcription factor MITF display changes in abundance (at the gene level) between melanoma subtypes; our approach could allow estimating the presence and abundance of the respective individual protein isoforms, and investigating how they vary across cancer subtypes, hence enabling a deeper understanding of cancer driving mechanisms.

PI's previous experience. The PI, as a statistician with solid programming skills and more than a decade of experience in mathematical modelling of biological data, represents **a rare interdisciplinary profile, ideal for the development of statistical methods for biological data**. Furthermore, he has already developed four Bioconductor R packages, and is familiar with various technical and conceptual aspects of this proposal: latent variables approaches, differential methods, Bayesian hierarchical modelling, and C++ coding.

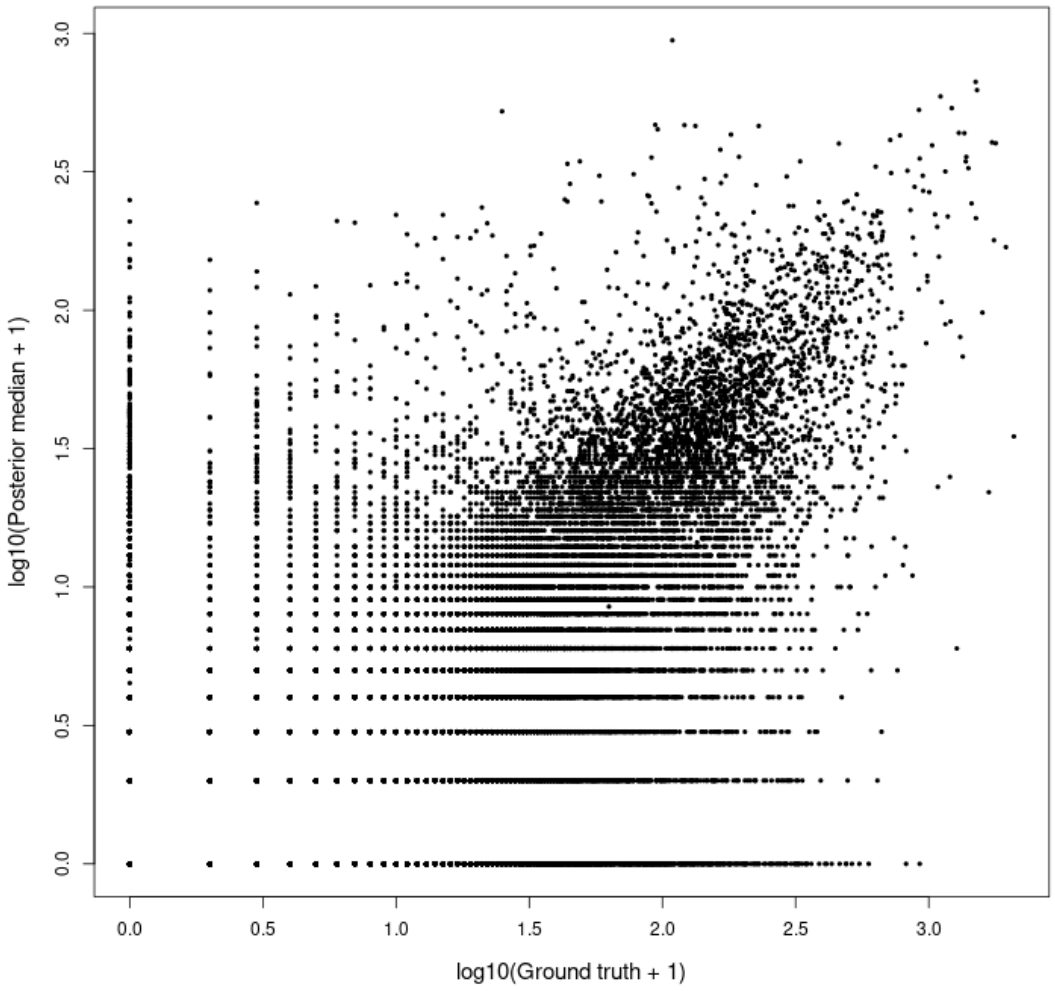

Figure 3: log-10 protein isoform abundances; estimates from our model (y axis) *vs.* ground truth (x axis).