Can Omics Biology Go Subjective because of Artificial Intelligence? A Comment on "Challenges and Opportunities for Bayesian Statistics in Proteomics" by Crook et al.

Thomas Burger

Can omics biology go subjective because of artificial intelligence? A comment on "Challenges and Opportunities for Bayesian Statistics in Proteomics" by Crook et al.

Thomas Burger*
Univ. Grenoble Alpes, CNRS, CEA, Inserm, Profi FR2048, Grenoble, France
* thomas.burger@cea.fr

Abstract: In their recent review,[1] Crook et al. diligently discuss the basics (and less basics) of Bayesian modeling, survey its various applications to proteomics and highlight its potential for the improvement of computational proteomic tools. Despite its interest and comprehensiveness on these aspects, the pitfalls and risks of Bayesian approaches are hardly introduced to proteomic investigators. Among them, one is sufficiently important to be brought to attention; namely, the possibility that priors introduced at an early stage of the computational investigations detrimentally influence the final statistical significance.

Keywords: Bayesian modelling, uncertainty representation, proteomics, mass spectrometry, statistical significance, artificial intelligence

Let us start with a toy example: As any p-value, the one associated to a protein differentially expressed (referred to as $Prot$) is counter-intuitive to interpret.[2] Thus, one may prefer the associated posterior error probability (PEP), reading:

$$\mathrm{PEP}_{Prot} = P(H0|Prot) = \frac{P(H0)}{P(Prot)} \cdot P(Prot|H0) = \frac{P(H0)}{P(Prot)} \cdot \mathrm{pvalue}_{Prot}$$

PEP_{Prot} quantifies the probability that $Prot$ is not differentially expressed, which directly interprets in terms of risk for the confirmatory wet-lab experiments. Unfortunately, to compute it from the p-value, one needs the probability of the null hypothesis $P(H0)$, i.e., the prior probability that a random protein is stable. If based on his faith in the project, an investigator (called Archibald) assumes 10% of the proteins to be changers while only 1% of them truly are, their PEPs are altered. Therefore, in Archibald's publication, the PEPs will not reflect the risk that confirmatory experiments discard the putative biomarkers. Naturally, here the error lies in Archibald's overconfidence, so that he should be the one to blame instead of the statistical theory he inappropriately used. Now, let us consider another investigator (Bianca) who finely estimates the prior to best guide her future confirmatory targeted proteomics experiments. She decides to rely on her knowledge of the pathways involved in the disease studied: Proteins of those have a lower probability of being stable than others. Doing so leads her to isolate a handful of putative biomarkers (which unsurprisingly, belongs to the pathways promoted by the prior). She will investigate them thanks to future grants, and meanwhile she prepares her manuscript about the discovery experiments. To compute the statistical significance values demanded by the journal, she naturally relies on the PEPs she computed and used. In other words, she does the same mistake as Archibald, which is to confuse how she deals with the risk/benefit balance along her investigations and the statistical significance of the very experiment she publishes. However, Archibald's mistake is easier to spot and arguably imputable to a data-dredging attempt, while on the contrary, Bianca's one is probably involuntary and certainly more difficult to identify. Finally, despite Bianca being apparently more scrupulous than Archibald, her approach is not immune to self-confirming biases (the pathways she promoted in the prior will provide significance values that will attract subsequent investigations).

While any statistical model can be used more or less correctly, we defend that Bayesian modeling, despite its power and versatility, can be a fertile ground for mistakes like Bianca's one. To understand why, a small historical detour about the interactions between biology, statistics and artificial intelligence (AI) is necessary.

In life sciences, statistics have long occupied a post-hoc quality control role, i.e., the objectivation that the discoveries claimed were not the result of luck. Because of the cost of wet-lab experiments or of clinical assays, preliminary statistical design has also thrived, but its role has mainly been to limit the risk of insufficiently significant statistics at the final stage. Therefore, until recently, biostatistics have essentially followed a goal of objectivity, in sharp contrast with the wet-lab investigations where the scientist's intuition, knowledge and skills, intrinsically subjective and personal, could act out. However, with the advent of omics biology, massive data are daily produced by high-throughput technologies, and computational biologists rely on multiple tools based on AI or statistics to correct batch effects, to profile patients or proteins (possibly undergoing imputations due to missing measurements), to visualize interactions with networks, and to build predictive models. In other words, this knowledge extraction process shares many formulae with significance computations, yet it is conceptually closer to the wet-lab investigation process: the way one makes the data speak along the course of a complex and iterative AI-aided process is partly subjective and does not share the objectivity aim of statistical significance.

The reason why statistical tools can be used both for significance assessment and for "subjective" inferences lies in the underlying probability theory. In the everyday language, a 0.5 probability has two meanings. It can either tell that success and failure, despite random, are equally frequent (e.g., head or tails game); or that in absence of sufficient information, the decision is random from the decision maker viewpoint, despite a possibly deterministic outcome (e.g., in which of my hands is the coin hidden?). This distinction between frequentist[3] and subjective[4] views has been almost as old as probability theory itself,[5] and while its implications have long been discussed in economy and game theory[6,7] or in AI,[8,9] they are scarcely so in omics biology.

As the natural tool to model the apparent random nature of the physical world, the probability of a given outcome should amount to its frequency of observation on the long run.[10] Such "frequency probabilities" have been instrumental to the development of statistical analysis, starting from demography[11] and progressively extending to a variety of application domains such as quality control,[12] insurance, etc. and of course, evidence-based medicine. Nowadays, a clinical trial is essentially a way to refine the frequencies of false positives and negatives associated to a given biomarker or drug. Indeed, both patients and policy makers would not accept that undesirable effects or false diagnostics occur at a higher frequency than originally claimed.

Another early application of probability theory has been gamble. However, gambling can also target unique events, for which frequencies of occurrences are either unavailable or unreliable (e.g., stock exchange crashes, football competitions, elections). To do so, partial information, intuition, and possibly emotions can be leveraged (e.g., accounting for the popularity of a football player, who could attract confident bets despite recent deceiving results). To this end, other theories have blossomed, where a probability is essentially a degree of belief in the decision maker's mind.[13,14] Such probabilities are called "subjective", "epistemic" or "Bayesian", as they are specific to a given state of knowledge (possibly incomplete or partially wrong).

Finally, probabilities have supported the development of two approaches: First, the observation-based modelling of a world submitted to randomness; and second, the digital encoding and combination of multiple state of knowledge for automated reasoning and AI. With Crook and his coauthors, we likely agree on several points: First, their review clearly makes the distinction between frequentist and Bayesian probabilities. Second, frequentist modelling, like the Bayesian one or any other one, comes

with limitations, so that hierarchizing them is pointless. Third, epistemic probabilities are more general than their frequentist counterpart, as it is always possible to build one's "subjective" viewpoint thanks to preliminary frequency estimations. Consequently, any epistemic framework is especially interesting for computational biologists, who rely on an expertise that is missing to the general-purpose data scientist to blend the results of an "objective" experiment (e.g., differential expression analysis) with more "subjective" background knowledge (e.g., the pathways most likely involved) as to propose new hypotheses, as ambitioned by Bianca. In this context, Bayesian methods are powerful and they deserve a better exposure in the proteomics community, a feature accomplished by Crook et al. Nevertheless, casting everything in a subjective formalism because of its larger conceptual genericity can be a double-edged sword: using a same language for AI-aided reasoning and for statistical significance analysis increases the risk of mating the latter with the former, as exemplified with Bianca's mistake. More precisely, when a protein is deemed as a putative biomarker because of its significant differential abundance, the associated significance value is classically understood in a frequentist context, for cultural reasons: Biologists have learnt to split the "subjective" investigation part from the "objective" (and essentially technical) significance computation one. Contrarily, for IA (including Bayesian) experts, there is a continuum from data modelling to decision making and their skill should express along this entire process, possibly using advanced tools (which description here would be off-topic) that make it possible to finely calibrate their priors as well as the subsequent significance values, to finally exhibit the best knowledge inference model possible. However, for computational biologist, it is probably more customary and safer to keep a clear albeit artificial distinction between the tools used for subjective AI-aided reasoning and those for statistical significance analysis.

To do so, specific tools are already available, as many researchers in AI have long identified that one weakness of Bayesian modelling lies in its single way of handling the various facets of uncertainty.[15] Although it is probably the reason of its success (any combination of subjective and frequentist probabilities yields a subjective probability, which can be subsequently combined, and so on) it is also its Achille's heel in an applicative context that facilitates confusions. To cope for this, multiple epistemic probability-like theories have developed in the last decades, where the [0;1] interval is used to encode different type of knowledge: Fuzzy set theory,[16] Possibility theory,[17] Dempster-Shafer theory,[18] etc. Essentially, they propose to make an explicit difference between the uncertainty resulting from randomness (classically captured by frequencies) and the one resulting from the investigator's subjectivity (possibly imprecise, partial, doubtful, etc.). In practice, the knowledge of equally frequent events in a head and tails game and the ignorance about which of the two hands holds the hidden coin are no longer encoded in the same way. Instead of focusing on the same gamble they lead to, one focuses on the different states of knowledge they correspond to, as to better incorporate them in an automated line of reasoning. In theory, relying on these frameworks should authorized more refined modeling. However, from the computational biologist's practical viewpoint, their application is hurdled by their complexity.

As relying on advanced Bayesian tools or on alternative frameworks for subjective inference are equally demanding in terms of theoretical skills, a third path has witnessed a growing popularity in computational biology, namely, empirical Bayes approaches.[19] Their central dogma is that priors should not be elicited by experts, as doing so would lead to too subjective biases that are incompatible with sound statistical significance. On the contrary, if the prior distributions are trained from the data, the final statistics can be expected to better generalize to new data. This is practically confirmed by the multiple empirical Bayes methods that are acknowledged in the state-of-the-art (e.g., Limma).[20] Despite, their incorrect use is not immune to opening a breach for subjective spoiling of statistical significance. Notably, if the prior parameters are estimated on the very same data as those subsequently processed, one increases the risk of reinforcing some data specificities (a.k.a., overfitting, another form of self-confirming bias). Let us illustrate this on the same PEP example, yet applied to Peptide-Spectrum Match (PSM) validation. Assuming a classical PSM score reading $S_{PSM} = -10 \cdot$

$\log_{10}(\text{pvalue}_{PSM})$, where pvalue_{PSM} is the p-value resulting from testing the match between the theoretical and empirical peak values in fragmentation spectra, it is possible to define the following PEP, which accounts for the peptide's length L:

$$\text{PEP}_{PSM} = P(mismatch|PSM, L) = \frac{P(mismatch|L)}{P(PSM|L)} \cdot \text{pvalue}_{PSM}$$

Doing so is insightful for two reasons: First, as the previous PEP, it easily interprets in terms of mismatch probability. Second, it purposely offers to account for the fact that random matches are not equally probable on shorter and longer amino acid sequences, for combinatory reasons. In an empirical Bayes setting, the ratio $P(missmatch|L)/P(PSM|L)$ can easily be "trained" from the data at hand: One essentially has to bin the amino acid sequences of the searched database according to their length. However, doing so can lead to unintentional significance dredging. Notably, if the searched database is too small, it may only contain a handful of very long peptides, with pairwise distinct lengths. Then, as each length L is specific to a peptide, $P(mismatch|L) = 0$, so that the PEP will be nil too, regardless of pvalue_{PSM}, as sometimes reported by MaxQuant's users.[21] In other words, whatever the quality of the match between the fragmentation spectrum and its theoretical counterpart, the PSM will be validated because the empirical prior had overfitted the data.

Finally, whatever the AI framework (i.e., classical Bayesian, empirical Bayesian, or even other subjective ones), the risk of mating the statistical significance with more subjective information exist and will increase, as inherent to the computational biology approach, and as a consequence of the ever-growing workflow complexity. Naturally, as in any scientific field, the evil is in the details, and the main safeguards we can think of are the scientists' rigor and integrity. Moreover, as sketched by Crook et al. in their review, Bayesian theory is rich and many advanced tools are available to refine the modelling. As for non-intentional dredging (i.e., Bianca's error), we propose two guidelines: First, whenever possible, keep separated computational workflows for biological knowledge inference and for statistical significance computation. Second, if it is not possible to separate them, provide an explicit formulation (both in mathematical and in everyday languages terms) of the prior knowledge incorporated in the AI-aided reasoning, as to subsequently foster a critic eye on possibly too optimistic biological conclusions.

References

[1] Crook, O. M., Chung, C. W., & Deane, C. M. (2022). Challenges and Opportunities for Bayesian Statistics in Proteomics. Journal of Proteome Research, 21(4), 849-864.

[2] Cohen, J. (1994). The earth is round (p<. 05). American Psychologist, 49(12), 997-1003.

[3] Venn, J. (1888). The logic of chance: an essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science, and to statistics. Macmillan.

[4] de Laplace, P. S. (1820). Théorie analytique des probabilités (Vol. 7). Courcier.

[5] de Moivre, A. (1718). The Doctrine of Chances: or, a method for calculating the probabilities of events in play.

[6] Jeffreys, H. (1939). The theory of probability (republished by Oxford University Press in 1998).

[7] Walley, P. (1991). Statistical reasoning with imprecise probabilities, Chapman & Hall.

[8] Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann.

[9] Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. (1995). Reasoning about knowledge. MIT press.

[10] Von Mises, R. (1981). Probability, statistics, and truth (2nd edition, English translation). Dover Books on Mathematics.

[11] Poisson, S. D. (1837). Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités. Bachelier.

[12] Student. (1908). The probable error of a mean. Biometrika, 6(1), 1-25.

[13] Cox, R. T. (1946). Probability, frequency and reasonable expectation. American journal of physics, 14(1), 1-13.

[14] De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In Annales de l'institut Henri Poincaré, 7(1)1-68.

[15] Dubois, D., Prade, H., & Smets, P. (1996). Representing partial ignorance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 26(3), 361-377.
[16] Zadeh, L. A (1965). Fuzzy sets. Information and Control, 8(3), 338–353.
[17] Dubois, D., Prade, H. (1988). Possibility Theory, New York Plenum.
[18] Shafer, G. (1976). A mathematical theory of evidence. Princeton university press.
[19] Carlin, B. P., & Louis, T. A. (2008). Bayesian methods for data analysis (3rd Edition). CRC Press.
[20] Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), e47-e47.
[21] Couté, Y., Bruley, C., & Burger, T. (2020). Beyond Target–Decoy Competition: Stable Validation of Peptide and Protein Identifications in Mass Spectrometry-Based Discovery Proteomics. Analytical Chemistry, 92(22), 14898-14906.

Acknowledgement: The author is grateful to Oliver Crook and to the anonymous reviewers for their constructive comments.

Funding: This work was supported by grants from the French National Research Agency: ProFI project (ANR-10-INBS-08), GRAL project (ANR-10-LABX49-01) and MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).