ELSEVIER

Understanding missing proteins: a functional perspective

Longjian Zhou[1], Limsoon Wong[2,3] and Wilson Wen Bin Goh[1,2,4]

Q1

[1] School of Pharmaceutical Science and Technology, Tianjin University, Tianjin 300072, PR China
[2] Department of Computer Science, National University of Singapore, 117417, Singapore
[3] Department of Pathology, National University of Singapore, 119074, Singapore
[4] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, 637551, Singapore

A missing protein (MP) is an unconfirmed genetic sequence for which a protein product is not yet detected. Currently, MPs are tiered based on supporting evidence mainly in the form of protein existence (PE) classification. As we discuss here, this definition is overly restrictive because proteins go missing in day-to-day proteomics as a result of low abundance, lack of sequence specificity, splice variants, and so on. Thus, we propose a broader functional classification of MPs that complements PE classification, discuss major causes, and examine three corresponding solution tiers: biological, technical, and informatics. We assert that informatics-driven solutions would have a major role in resolving the MP problem (MPP).

Introduction

The publication of the first draft maps of the human proteome [1,2] (~19 000 proteins) has invigorated interest in MS-based proteomics as a high-throughput analytical platform, with critical clinical implications. Given current hardware advances, we can obtain unprecedented proteome coverage. Yet, these draft proteome projects do not reflect day-to-day proteomics. Unlike genomics, where nucleotide sequence and quantitation information are assayed as a single information unit, sequence and quantitation are assayed separately and indirectly in proteomics. MS-based proteomics only measures the mass:charge (m/z) ion ratios, retention times, and ion intensities. Reverse-engineering these into sequence and quantitation across a dizzying array of protein moieties (e.g., the primary protein sequence, splice variants, and post-translational modified forms) in a single sample is difficult. Moreover, in data-dependent acquisition (DDA) setups, spectra are acquired inconsistently because of semistochastic preselection. As different spectra (corresponding to different peptides) are identified across different runs, protein identification and/or quantitation (protein quantitation is inferred from detected unambiguous constituent peptides) becomes unstable. Consequently, a typical proteomics experiment can only yield with high inconsistency a small fraction of the proteome. In other words, given a single proteomics run, a large fraction of the proteome is 'missing'.

Q2

In contemporary parlance, the term 'missing protein' (MP) takes on a narrower definition. The term 'MP' was first coined by Paik *et al.* [3] while delineating the goals of the chromosome-centric Human Proteome Project, cHPP, which aims to detect via MS at least one protein for each known human gene sequence. Any gene sequence whose respective protein has never been observed is an MP. Alongside various initiatives, such as GPMDB [4], PeptideAtlas [5] and neXtProt [6], the goal is to establish a genome–proteome bridge.

To systematize efforts, MPs are classified according to the five neXtProt Protein Existence (PE) tiers [7] (Table 1). PE1s are not MPs because they have protein-level support. Subsequent tiers lack protein evidence. PE2–4s only have transcript, homological, or prediction support, respectively. PE5s are highly dubious entries, but they represent a relatively large proportion of MPs; thus, efforts are underway to confirm their veracity [8]. Interestingly, many PE5 s are membrane-bound proteins, which are historically difficult to isolate and detect [8]. PE classification holds some unexpected

Corresponding authors: Wong, L. (wongls@comp.nus.edu.sg), Goh, W.W.B. (goh.informatics@gmail.com)

1359-6446/© 2017 Elsevier Ltd. All rights reserved.
https://doi.org/10.1016/j.drudis.2017.11.011

TABLE 1

Q8 **Protein existence (PE) classification of proteins**[a]

PE tier	Inclusion criteria	Estimated percentage of proteome[b]	Notes
1	Evidence at proteome level	~82.0% (16 518)	At least two unique nonoverlapped peptides at least nine amino acid-residues long
2	Evidence at transcript level only	~11.5% (2290)	The transcript must be confidently detected, but with no corresponding protein evidence
3	Homology inference only	~3.0% (565)	Inferred homologs without protein or transcript support
4	Predicted	~0.5% (94)	Predicted coding sequence, without homology, transcript, or protein support
5	Dubious	~3.0% (588)	The sequence might not fully meet the criteria for a predicted coding sequence; uncertainty over the veracity of the coding sequence (i.e., we do not know the sequence is correct); some studies do not consider PE5 as MPs

[a] Current rules, as of 2017.

[b] As of 2016 against a total of 20 055 proteins (estimated numbers).

evolutionary significance: PE3s and 4s are young, spreading from nonhomology chromosomal regions and exhibiting higher sequence divergence rates [9]. Unsurprisingly, given that many are derived from duplication events, they are sequentially similar and often tissue specific and, thus, difficult to detect (only unambiguous information is used for protein detection).

MPs constitute only approximately 18% (3610 proteins) of the known gene sequences (Table 1). It is a small proportion, but nonetheless an important uncharted territory, comprising many unique genes with clinical potential (e.g., biomarkers or drug targets, provided their respective proteins are targetable). Many MPs are low abundance [10] or membrane bound [8], falling outside the capabilities of conventional MS.

MPs are a solvable problem from the conventional viewpoint [9], given that 82% of human proteins are detectable. However, we stated above that most of the proteome is 'missing', and many proteins are inconsistently detectable in a typical proteomics run. To us, an MP can be one of the following: sequence is known but hard to detect; sequence is known but never detected in MS; or sequence is not known but evidence exists for it, such as via gene prediction or in raw spectra. The MPP (i.e., the difficulty in detecting certain proteins despite transcript or theoretical evidence) should more rightfully be considered a narrow manifestation of the more-general coverage (i.e., the inability to survey the entire proteome) and consistency (the inability to consistently detect a protein) problems [11–13].

Thus, a broader perspective is necessary. Here, we generalize the notion of an MP, explore the underlying causes and extent of the problem (including how it leads to failures in functional analysis), classify and evaluate solutions (particularly for correct interpretation of received data), and assert that informatics-driven solutions are the most practical way forward.

Generalizing the MPP

Most proteins are observable to some degree. The more pertinent question should be whether they are consistently and reliably detectable for practical purposes (e.g., clinical diagnostics or functional proteomics).

Of 20 055 human proteins, approximately 18% (3610 proteins) are MPs [14]. Resolving these is nontrivial, requiring several collaborating laboratories [7] and dedicated databases [15]. Although resource heavy, this is a problem that is eventually solvable (perhaps with the exception of PE5). However, this is also a limited perspective. In routine proteomics, proteins often go 'missing', because the protein poses technical challenge for MS, or the spectra search was not configured correctly or optimally [e.g., splice variants, post-translational modifications (PTMs), and inappropriate tolerance ranges for peptide-spectra matching (PSM)].

More often, and with serious consequences, a protein is partially observed in some samples but not in others, creating serious statistical and functional analysis issues. In severe cases, only 10% of detected proteins are common to all samples, making comparative analysis impossible [16–18]. It is estimated that, in a typical experiment, only 30% of the spectra are decipherable (i.e., mappable to peptides); thus, the majority of the proteome is technically invisible (i.e., missing) [19,20]. Although reduced coverage is an obvious consequence, being unable to survey sufficiently large segments of spectra also leads to quantitation instability [19]. To us, this generalization of MPP has far-reaching consequences: even if we resolved the remaining approximately 18%, that proteins are difficult to observe routinely or cannot be quantitated properly does not make proteomics a practical technology.

Why do proteins go missing?

Currently, an MP is a gene sequence for which a protein product has not been observed. In our broader definition, MPs also include those that are missed in routine experimentation. These stem from diverse causes broadly classifiable into biological, technical, and informatics. Here, we highlight three pertinent issues from each category (Fig. 1).

Biological issues stem from inherent protein properties. Low-abundance proteins are harder to detect if below the detection limit of the instrument used. Yet, many low-abundance proteins are ostensibly important; (e.g., cell-cycle-related proteins are known to constitute the cancer-proliferation signature and, thus, are important for diagnostics) [21,22], but are poorly detected by proteomics [10].

Whereas low-abundance issues are well known, sequence ambiguity is less well appreciated: Webb-Robertson *et al.* demonstrated that the lower the number of unique peptides in a protein (and, therefore, the more ambiguous), the more likely the protein is to go missing (because of a higher chance that its small number of unique peptides might not be observed in a proteomics screen) [23]. Same-family proteins are especially like to have ambiguity issues.

Please cite this article in press as: Zhou, L. et al. Understanding missing proteins: a functional perspective, Drug Discov Today (2017), https://doi.org/10.1016/j.drudis.2017.11.011