Libraries of Specific Assays Covering Whole Proteomes: From Yeast to Man

N. Leigh Anderson*

Although only approximately 200 protein analytes, equivalent to 1% of the human proteome, are currently measured in plasma or serum for clinical purposes, there is an enormous unmet need for new or improved clinical diagnostic tests in many disease areas. Substantial biomarker discovery efforts over the last 5 years have generated hundreds—in some cases, thousands—of candidate protein markers in many of these areas, and recent attention has focused on how we can rigorously evaluate such candidates in large sample sets to identify the handful that have real clinical value. Faced with the need to sift many candidates to find a few useful diagnostics, one immediately confronts a major barrier in the lack of reliable assays capable of detecting most human proteins in the relevant clinical specimens (typically serum or plasma). It is estimated that ELISA kits exist in some form for perhaps 2000 human proteins, but the quality of most of these is unlikely to approach that of the 200 highly-optimized clinical protein assays, which themselves have taken years to develop and occasionally still have marked specificity uncertainties. The cost and time required to build really good immunoassays for thousands of candidate protein biomarkers is difficult to estimate, but could run into hundreds of millions of dollars over a decade or more. A further requirement for testing many candidates in the very limited sample volumes obtained in well-structured trials (i.e., assay multiplexing) further constrains most immunoassay platforms; hence the need for a more systematic solution to the problem of protein measurement if we are to see rapid progress in the protein biomarker field.

Picotti et al., in Ruedi Aebersold’s laboratory at ETH in Zurich, recently demonstrated an important and encouraging advance in this direction by developing tools (1) to measure all the proteins in a well-characterized model organism: yeast. In principle, the complete genome sequence provides all the information needed to specify amino acid sequences unique to each of the approximately 6000 yeast proteins. Given a quantitative measurement method capable of recognizing such sequences, one would, in theory, be able to generate an assay for any yeast protein with genome-style (i.e., sequence-based) absolute specificity. Mass spectrometry provides such a tool in the form of the selected reaction monitoring (SRM) assay of proteotypic peptides present after complete digestion of a yeast sample by a protease such as trypsin. In this context, a proteotypic peptide is one whose sequence is unique to the target protein (and hence whose molar amount indicates the molar amount of the protein in the starting sample) and that is well detected in the relevant mass spectrometer system. In SRM assays, a 2-stage mass-selection system first selects the known mass of the peptide (calculated exactly from the amino acid sequence), then fragments the peptide (generally through collision with gas atoms), and finally selects 1 specific fragment of known sequence, which is counted at a detector. The pair of peptide and fragment masses (or, more correctly, mass-to-charge ratios) is called a “transition,” and its measurement provides great specificity that can be further increased by measuring several transitions per peptide. Hundreds to thousands of these transitions can be measured in a repeating series during chromatographic separation of a complex peptide digest. By supplementing samples with known amounts of stable isotope–labeled synthetic versions of target peptides, separately quantified by parallel but different SRM transitions, true internal standardization can be achieved. The SRM method has been widely used on nonpeptide analytes such as drugs and metabolites for many years, and is used increasingly in clinical laboratories for the measurement of immunosuppressants, vitamin D, steroid hormones, and aberrant metabolite patterns during newborn screening.

Picotti and colleagues have taken this principle and made its application tractable on a proteome-wide scale by showing how optimized parameters for SRM assays of proteotypic peptides can be generated rapidly using libraries of crude synthetic peptides analyzed approximately 100 at a time. Critical assay parameters include choice of highest-intensity fragment ions, peptide elution time during chromatography separation, and choice of secondary fragment ions useful for specificity confirmation. This approach, of manufacturing critical peptide experimental data using relatively...
cheap crude peptides, can provide coverage of all the proteotypic tryptic peptides of yeast, and thus extends the scope of an earlier database of yeast SRM assays based on proteomics methods of lower analytical sensitivity, covering only approximately 1500 proteins observable by untargeted so-called shotgun methods (2). The group has shown that SRM detection in the yeast system can measure proteins ranging from >1 000 000 down to approximately 50 copies per cell, and after resolution into peptide fractions by off-gel isoelectric focusing, down to single-digit copies per cell, thus providing coverage of the full dynamic range expected for functional yeast proteins (3).

Can this approach be applied directly to solve the human protein measurement problem? In terms of SRM assay design, the answer is yes in principle, with practice soon to follow. The approximately 20 000 human proteins represent a manageable step up from approximately 6000 in yeast, and plans have already been laid to create a large database of human SRM assays based on experimental data from more than 100 000 crude synthetic human sequence peptides (Robert Moritz and Ruedi Aebersold, personal communications, July 9 and 14, 2010, respectively). For humans, a further step is being planned, in which each human protein (or selected parts of large proteins) is being expressed in human cells (Robert Moritz, Lee Hood, and Wei-Wu He, personal communications, July 9, 2010) or in human in vitro expression systems (Michael MacCoss, personal communication, July 9, 2010), then digested to yield its characteristic tryptic peptides in the correct equimolar mixture based on recovery. Analysis of these digests provides additional certainty as to which peptides have the highest relative detection efficiency by SRM, and covers all the peptides instead of the 20% proteotypic subset typically chosen for synthesis. As these projects proceed, we should be able to expect a comprehensive database of SRM assays covering all human proteins, or at least 1 representative protein from each gene, within the next 2 to 3 years. This expectation, bolstered by the availability of a database of theoretical human proteotypic peptides digested in silico, makes the sequences of a majority of peptides that will ultimately be used in clinical SRM assays obvious from a patent point of view, and thus is likely to establish the chemical structures of these fundamental analytes, and potential diagnostic markers, in the public domain.

The remaining challenge in applying SRM assays to real clinical samples is analytical sensitivity. SRM alone can cover the approximately 20 000-fold range between an approximately 1 000 000- and a 50-copy-cell protein in yeast, very similar to the limited range it covers in digests of human plasma (i.e., serum albumin down to proteins at about 1 mg/L). Complete coverage of known proteins in human plasma, however, involves spanning the much larger approximately $10^{10}$-fold range in concentration between albumin, on the one hand, and cytokines or high-sensitivity cardiac markers on the other. It has become clear that additional specific enrichment or fractionation steps, in combination with direct sensitivity improvements in MS instruments, are needed to provide the extra 300 000-fold analytical sensitivity required to cover the whole range of existing protein diagnostics—and these advances appear to be close (4), if not already in hand for routine use. Once completed, the atlas of human SRM assays, with associated analytical platforms, opens the door to much more powerful protein diagnostics, including multianalyte panels, while reinforcing the linkage between clinical laboratory medicine and basic medical research.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures of Potential Conflicts of Interest: No authors declared any potential conflicts of interest.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References