The Proteomic Search for Diagnostic Biomarkers

LOST IN TRANSLATION?*

Werner Zolg‡

The search for biomarkers based on proteomic discovery strategies is being conducted in different scientific constellations. One major group is organized around government-sponsored programs involving a network of academic institutions focusing on biomarkers systematically. An example of such a multicenter initiative for the United States is the "Directors Challenge Program," which aims to characterize the molecular basis and related changes in protein expression of major cancers (www.cancerdiagnosis.nci.nhi.gov/challenge). Simultaneously there are major research and development programs in the pharmaceutical industry also focused on the identification of biomarkers but for different reasons (1). Various alliances and partnerships have also been formed between academic institutions and between industry and academia. Focusing throughout on *diagnostic* applications of biomarkers and ignoring the even larger efforts to identify proteins that can serve as drug targets, the goal of industry-driven efforts is not primarily the identification of diagnostically relevant proteins itself, but clear economic imperatives call for the eventual introduction of those novel biomarkers into the market. So far, financial analysts estimate that close to one billion dollars have been spent by or invested into companies with business models based on proteomic strategies for the identification of drug targets and/or biomarkers, including diagnostic biomarkers.

However, there are few, if any, new biomarkers that have resulted in a commercial product by completing the proteomic process chain of identification, validation in clinical trials, and approval by regulatory agencies. On the contrary, the number of new protein targets approved by the registration agencies has been declining over the past decade, a trend not reversed or even stopped by any of the ongoing activities in proteomics. Moreover a survey of commercially available products used as diagnostic tools reveals that an amazingly small number of proteins in the human proteome have been identified as diagnostically relevant targets, but instead entire assay families have been developed that target the very same proteins in different configurations (2). Furthermore the scientific community as well as the general public has been led to believe that major diagnostic breakthroughs have already been achieved as a result of activities in proteomics and that consequently a series of novel diagnostic biomarkers are in place to fill the diagnostic gaps in a reliable and beneficial way. This misleading information policy has already proven to be counterproductive because expectations have been raised that have ultimately not held up to scrutiny. There have been high investments on one side but a low success rate on the other. Given this discrepancy, one feels challenged to start analyzing some of the potential hurdles in the processes leading from the discovery phase to the validation phase to product development. The rationale for these sequential steps has already been outlined (1, 3, 4), but several aspects should be revisited from different angles. Some of the considerations below were presented at the 21st Asilomar Conference on Mass Spectrometry in October 2005.

CONSIDERATION 1: MAKING SELECTIONS

Even before initiating the discovery phase, the first step in the process chain of creating new diagnostic content is to make critical decisions that will directly impact on the outcome of the identification process. The very selection of the discovery samples and their degree of characterization, down to the standard operation procedures (SOPs)[1] of how the samples were acquired and stored, can be decisive for success or failure. The principle "garbage in, garbage out" will undoubtedly be confirmed when, in the end, inconsistent differences in the protein patterns must be interpreted.

Sample Types—Leaving the serum proteome aside (see "Consideration 3"), the hypothetical discovery program (*e.g.* an oncology marker) will likely be based on tissue. One should take advantage of the speculative fact that the potential marker candidates will be present at a higher concentration in the compartment in which the disease process actually takes place (tissue) than after dilution in peripheral blood. By selecting tissue as the discovery material, one must inevitably choose between cultured cells or specimen directly obtained from patients. There are advantages to either option, but if the direct tissue comparison is chosen, there are many pitfalls. The degree of tissue characterization is critical to the identi-

From Centralized Diagnostics, Roche Diagnostic GmbH, Nonnenwald 2, D-82377 Penzberg, Germany

Received, January 10, 2006, and in revised form, February 3, 2006
Published, MCP Papers in Press, March 17, 2006, DOI 10.1074/mcp.R600001-MCP200

[1] The abbreviations used are: SOP, standard operation procedure; MAR, multiple affinity removal system; PRS, protein reference standard; CRC, colorectal cancer; RA, rheumatoid arthritis; OA, osteoarthritis; ROC, receiver-operator characteristic; AUC, area under the curve; CI, confidence interval.

fication in the subsequent analytical steps. It is of utmost importance that the specimens used in the diseased cohort are not simply classified as "tumor" (together with possibly the stage of the tumor) but that a detailed histopathological assessment of the distribution of cell types (*e.g.* tumor cells, necrotic cells, and stroma) in the diseased specimens is carried out. This distribution should be as uniform as possible in all discovery samples because otherwise normalizing the signal intensities obtained in later fractionation steps becomes very difficult, opening the possibility of misclassification or missing interesting marker candidates altogether.

Sample Numbers—The number of samples being placed in the "diseased" and "healthy control" groups to be compared with a variety of analytical approaches remains a matter of discussion. If the number is for practical reasons (time lines and resources) and kept small (<10 per group), then the observed differences between the two sets of specimen are in danger of being overinterpreted when extrapolated to generalized cohorts. The high amount of interindividual differences should not be underestimated. Low sample numbers make correctly identifying those differences increasingly difficult. In early proteomic discovery programs we had to run second and third discovery rounds to complement the results of the first round to confirm or reject selections based on the initially low sample numbers. Ideally the sample number analyzed should not only allow one to state the presence or absence of a given protein but also to follow *trends* in the distribution of proteins among the individuals in the collective. The question of throughput will eventually govern the decision making concerning sample numbers. As a compromise, we feel that a minimum of 15 samples in the discovery phase is necessary to get a reasonable representative selection basis for marker candidates. We have completed the discovery phase for colorectal cancer markers using sample collectives of this size (5), creating protein lists that allow for comparisons leading to some 40 marker candidates to be carried forward into the validation phase. The size of the validation collectives must be quite different (see "Consideration 6"). As the number of samples in the discovery phase drops, the risk of being misled by "one-hit wonders" increases. These false identifications will use up resources in the subsequent validation process without increasing the chance of obtaining verifiable results in representative sample cohorts.

To circumvent the dreaded game of numbers, it is often suggested to pool the samples, *i.e.* to physically combine several of the extracts to create fewer samples to be put through the entire analytical process. This strategy may seem tempting because substantial conservation of resources and manpower can be achieved. However, in our experience, pooling of samples inevitably leads to loss of information. The distribution of proteins is averaged by the very pooling process with the prospect that individual proteins are pushed under the detection limit by one member of the pooling cohort not expressing the protein in question. At any rate, somewhere in the selection process, the individual spectrum of proteins has to be established. Thus the pooling process just shifts the workload to a later point in the process chain, and really good arguments have to be found to deliberately increase the complexity of the data sets by pooling.

CONSIDERATION 2: SETTING PRIORITIES

In the overall perception of many, proteomics is strongly associated with the discovery phase, the first step in the process chain to create diagnostic content. Proteomics is reduced to identifying proteins as being "different" in various states of a biological system classified as diseased and "healthy." As a result, there is a clear tendency to overrate the discovery phase if the ultimate goal of programs in proteomics is limited to the context of *in vitro* diagnostics, namely making a novel validated biomarker available with a clearly defined position within a diagnostic chain. Consequently the absolute number of discovery hits compiled on an Excel sheet cannot be considered the primary goal of proteomic approaches in the *in vitro* diagnostic setting. On the contrary, the true challenge will be the assessment of the selected candidates within carefully planned validation schemes. The subsequent validation phase, not the discovery activities, will reveal whether there is something interesting and valuable with a prospect to be able to fill the existing sensitivity and/or specificity gaps.

The validation strategies will vary depending on the intended positioning of the marker. A *screening marker* has to fulfill completely different performance criteria than, for example, a *monitoring marker* (a diagnostic situation in which the disease is already known) or an *efficacy marker* modulated by the drug being administered. To increase the likelihood of success in industry-driven programs, marker candidates should be selected with the potential to complement already existing product lines aiming to increase the performance of the overall package. Building upon existent product lines is an economic necessity because novel replacements are bound to face high "entry barriers" into the market. In addition, there is little guidance from regulatory agencies as to the specific benchmark(s) against which the single novel marker must be compared. A real challenge will be gaining approval for linking novel and known markers together using mathematical models ("disease algorithms") based on multiparametric testing. Given these complexities, I am intrigued by many of the simplistic approaches still practiced by some proponents in the field. With the ultimate goal of a discovery phase, they advocate the identification of a *single* marker that will supposedly describe complex diseases in the multitude of diagnostic settings. This "golden bullet approach" is, at best, an illusion and should, based on everything we have learned in years past, be considered an unrealistic scenario. It is of little help for the overall goal that these scenarios are revived at opportunistic times in presentations and press releases.

CONSIDERATION 3: FACING THE INEVITABLE: THE COMPLEXITY OF THE SERUM/PLASMA PROTEOME

When starting a discovery program directly in serum, one is immediately confronted with an important dilemma that is now even discussed in textbooks: the concentration range of serum proteins spans some 12 orders of magnitude, and the analytical approaches, regardless of their type and nature, cover about 3 orders of magnitude (2). There is simply no way that, without reducing the complexity of the protein mixture within the human serum proteome, mass spectrometry-based approaches can identify analytes in the concentration range of common tumor markers (ng/ml range). A model calculation can be found elsewhere (1). Consequently one must remove highly abundant serum proteins to increase the sensitivity of the subsequent analytical procedures. Internally we call this labor-intensive process "climbing down the Anderson Ladder" (2). A quick inspection reveals the nature of the highly abundant proteins and their concentration. Six of the highly abundant proteins (albumin, IgG, IgA, transferrin, antitrypsin, and haptoglobin) can be removed by a commercially available immunoadsorber based on polyclonal antibodies (Multiple Affinity Removal System; MAR, Agilent Technologies, Waldbronn, Germany).

In our efforts to make *reproducibility* the highest ranked criterion for success, we control the effectiveness of MAR column depletion by using highly calibrated automated ELISAs to determine the concentration of the six proteins to be removed in the flow-through. With this monitoring method, we can quantitate the "creeping-through phenomenon" after multiple runs and learn about the reduction of complexity that directly impacts the next fractionation steps. Although the seven most abundant proteins amount to about 97% of all the serum proteins, it is known that some additional 30 proteins are present in the greater than μg/ml range. Additional commercial tools that met our internal quality specifications were needed to remove members of this subfraction of the serum proteome. Because these were not available to us, we opted for a sort of "brute force" approach; we started the development of a series of immunoadsorbers, making use of one of our core capabilities: raising monoclonal antibodies in gram amounts. In addition, we can rely on an established internal work flow to monitor the binding characteristics of any monoclonal antibody with regard to the later test requirements by affinity measurements (*e.g.* pH stability and salt conditions). Again each of the new adsorbers was characterized by monitoring the binding capacity and effectiveness using automated ELISAs for each protein in question. In addition, we carefully monitored the *nonspecific* binding of proteins to each of the new adsorbers by spiking-in mixtures of proteins as tracers (*e.g.* selected tumor markers)[2] while monitoring their recovery by automated Elecsys[3] assays (Roche Diagnostics GmbH, Mannheim, Germany). So far we have completed the development of nine additional immunoadsorbers with more to be completed in fixed time intervals. At least conceptually, it would be desirable to remove, for example, the top 30 serum proteins with proven SOPs to increase the analytical sensitivity of MS identification processes. Achieving such a goal ultimately depends on the future allocation of resources and time to proteomic discovery programs at Roche. At least, for the time being, highly depleted serum (total serum minus defined highly abundant proteins) is available to jump start the comparison of the protein patterns derived from multiple disease states directly in plasma.

However, any extent of protein removal is only the first step toward reducing the complexity of the protein mixtures in serum. This first step must be embedded in an overall fractionation scheme that involves coherently linking known fractionation principles and making sure that the buffers are compatible in each sequential step to arrive at defined subfractions of the plasma proteome. The more the low abundance proteins (containing the assumed biomarkers) are enriched within the pool of remaining high abundance proteins, the better the chances are of reliably identifying those biomarkers with the subsequent analytical procedures. Tilting the balance in favor of the low abundance proteins is limited by the number of fractions one is prepared to analyze per proteome. Despite all the automation steps and the options for upscaling that can be realized in the sequential fractionation and analytic processes, it comes down to the question of throughput and thus time lines to results.

CONSIDERATION 4: STANDARDIZING ANALYTICAL METHODS

Regardless of which elaborate scheme is implemented to fractionate plasma before peptide fingerprints or sequences are obtained, it is critically important to control the fractionation status and thus the *reproducibility* of the procedures used. This is less critical in processes in which different protein mixtures are labeled with selected reporter molecules and subsequently mixed and combined through the entire fractionation cascade. The deviations, however large and erratic, will affect both protein sets identically. However, if one wants to compare the protein composition of two (or more) parallel serum proteomes relying on the peptide annotations after mass spectrometry, highly standardized fractions are an absolute prerequisite for a meaningful comparison of the final results. We implement a rigid scheme of standardization: depleted plasma (see "Consideration 3") is spiked with an internal protein reference standard (PRS) previously prepared in multiple identical aliquots. Thus, the same reference points are available throughout for all our internal programs in proteomics independently of the indication area. PRS consists of a mixture of six non-human proteins with different isoelectric points and molecular weights mixed together at differing molarities. Tracing preidentified peptides of the spiked-in standard proteins by two-dimensional LC ESI-MS/MS gives an

[2] S. Palme, personal communication.
[3] Elecsys is a trade name of Roche.

accurate picture of the reproducibility of the methods used.[4] One might question why a similar standard has not been propagated by the proteomics community from the inception of discovery activities. If generally accepted internal protein standards were available and reference to their uses was mandatory in publications, all discussions regarding the sensitivity of different analytical methods would finally be grounded on rationality.

The most critical question after completing the comparison of two protein profiles remains "when can a protein be considered a marker candidate?" The answer is relatively straightforward if a protein is present in the majority of the diseased sample sets (preferably in all samples of the set) and absent in all or the majority of the healthy controls. The opposite situation, often the most revealing, is not attractive to the diagnostic industry. A far more challenging situation arises if there are only quantitative differences among the sample sets. One must either establish quantitative MS approaches or use isotope labels (6). The first option is still subject to experimental evaluation and validation in many proteomics laboratories. The latter may be tempting, but a series of validation experiments must be carried out to make reliable estimates about the possible losses in overall sensitivity and the reproducibility of the isotope ratios.

CONSIDERATION 5: VALIDATION STRATEGIES

After completing the discovery phase, the comparison of the proteins identified in the two proteomes will eventually lead to a consolidated list of marker candidates. The next phase is to develop a validated biomarker, and this phase (also known as the validation or prototype phase (1)) is more challenging than the discovery phase for several reasons. As evidenced by the number of publications, discovery programs attract far more interest than serious validation programs. A multitude of factors might account for this; for example, there is a high attrition rate as marker candidates undergo the process of becoming validated markers, and there is limited interest in communicating negative results. The validation process is definitely more costly and labor-intensive than any comparable discovery program we have ever conducted, the time lines being one of the main deterrents. Yet because the value of a marker in a diagnostic chain is solely determined by its validation results, we shift resources and manpower from the discovery phase to the validation phase as soon as a set of marker candidates (three to five) per indication area is available. The bottleneck of validation is not overcome by filling the discovery pipeline even further. It is also part of the program leadership to convince the team members that the perceived "unglamorous" aspects of validation are more than compensated by the "value-creating element" of the validation processes. This is especially true if large validation panels must be manually worked up in a repetitive way (see "Consideration 6"). In my opinion, the often cited success factor "innovation" in the context of biomarkers has to take hold in the validation phase to a far greater extent than in the discovery phase. Selecting epitopes of target molecules to obtain antibodies with an acceptable specificity or getting rid of hook effects in the ELISA prototype design can, in fact, be more challenging than identifying differences in the protein composition of two discovery sample sets. Public discourse, therefore, should also support the notion that scrupulous validation work should receive as much recognition as the initial discovery activities.

It is our belief that the biomarker should ultimately be detected in serum/plasma using immunological detection (ELISA) formats. Let us assume that the marker candidate was identified by comparing sets of tumor and healthy control tissues. Next selected immunogenic peptides must be synthesized, and full-length recombinant proteins (expressed in different pro- and eukaryotic expression systems) must be purified. In some cases, cDNA (for DNA immunization) complements the family of immunogens used to raise antibodies. For a first screen, we opt for throughout and use polyclonal antibodies raised in rabbits. Because the immunization will usually be running for 100 days before the final bleeding, several of the above approaches will have to be run in parallel increasing the number of animals per marker candidate to up to 12–15. The more economical sequential approach, using one immunogen after another, would lead to unreasonably staggered time lines. Once the antibody batches are available and the IgG fraction has been precipitated (we consider negative immunoadsorption as too labor-intensive at this validation stage), the antibodies will be used in extensive prevalidation schemes. In one of the early verification steps, the first round antibodies are used in immunostaining of tissue sections previously used in the discovery phase. Next Western blots with lysates from the same tissues used in the discovery phase will confirm the modulation in the expression level of the antigens in question. In addition, the Western blots using lysates from a variety of tissues are a first check for the specificity of the antibodies raised. An example from the discovery and prevalidation phase of the colorectal cancer program can be found elsewhere (6).

Unfortunately this first antibody check is confined in most cases to tissue itself because Western blots with serum require the tentative marker to be present in the μg/ml range. Therefore, the decisive question of the process (*i.e.* whether the antibody will recognize the presumed biomarker directly in serum/plasma) often remains unanswered until a prototype ELISA is established. This assay format allows for, in an optimized configuration, the detection of analytes as low as 20 pg/ml. By no means will all the antibodies raised in the first round be adequate to serve as capture entities in an ELISA format. Consequently a second or even third round of immunization is necessary for detection of the sought-after analyte in serum. This approach requires substantial resources in

[4] M. Thierolf and G. Pestlin, personal communication.

peptide synthesis, protein chemistry, and animal facilities together with the commitment to follow up with each candidate for months before they can be eliminated or carried forward to the next validation steps. The absence of complete antibody libraries against all expressed human genes is a situation that is about to be remedied by the tremendously important initiative of establishing a human protein atlas for normal and cancer tissues (7). Clearly this laborious sequence of identifying marker candidates followed by raising multiple antibodies is the most challenging step with a very unfavorable input-output ratio.

Alternatively to establish whether candidates identified in areas of presumably high concentration (synovial fluid in rheumatoid arthritis, cerebral spinal fluid in central nervous system diseases, or tumor tissue in oncology) can be traced in serum without first raising antibodies, we used quantitative multiple reaction monitoring mass spectrometry based on spiking in a set of ^{13}C-labeled peptides, derived from the protein in question, as reference points (8). Although multiple reaction monitoring was successful in identifying and quantifying certain sets of proteins directly in serum, we only use this technology for specialized questions and not in routine validation processes mainly because the technology is labor-intensive, the labeled peptides are a definite cost factor, and for some proteins we simply do not reach the analytical sensitivity to trace them directly in serum.

CONSIDERATION 6: SAMPLE BANKS OR PUTTING THE CANDIDATES TO THE TEST

Once the biomarker candidates can be detected in serum using ELISA prototypes, only the *first* qualifier hurdle in the validation process has been passed. Next we designed a two-step testing approach: assuming the marker is intended to be part of a screening scenario we use optimized ELISA prototypes to measure 50 highly characterized samples from diseased individuals and 50 samples from healthy blood donors. This limited "black and white panel" (panel A) is the *second* qualifier in the process of transforming a marker candidate to a validated marker. It should be noted that this particular qualifier principle is restricted to screening scenarios only and that panel A is totally inadequate for classifying markers in monitoring scenarios for example. If a candidate does not correctly classify a preset fraction of the two sample collectives (see "Consideration 7"), then its diagnostic value as a stand-alone marker is limited (or non-existing), and lower priority should be given to this candidate in the follow-up testing. However, eliminating markers solely based on the results of panel A is inherently risky: markers with a low univariate discrimination power might in fact become part of a disease algorithm based on mathematical models applied for multivariate analysis. Those markers might provide additional information to an existing set of markers, thus improving the "disease algorithm."

Following the ranking based on panel A, a *third* qualifier is applied. A serum/plasma panel of up to 1,500 samples (panel B) is challenged with the candidate in question. To be included into panel B, the samples must be available as both serum and as EDTA-anticoagulated plasma with a minimum volume of 5 ml so that each potential marker for a given disease can be evaluated over the entire validation phase with the same samples. The sample collection itself and the storage of suitable aliquots follow rigorous identical SOPs that are mandatory for all participating collection centers to avoid known variations in measuring the analyte concentration similar to those recently described (9). Each sample is accompanied by extensive case report forms containing all information (up to 300 data points). This allows for the eventual formation of subgroups as needed according to previous medication, co-morbidities, disease duration, and other classifiers.

Using colorectal cancer (CRC) as an example, CRC panel B consists of seven subgroups I-VII (the number of samples within each subgroup is given in parentheses). Subgroup I (250) contains CRC cancers with colonoscopy results split into four different cancer stages according to the incidence available for the United States in the 50–59-year population. Subgroup II (320) contains "other cancers" with a minimum of 25 cancers per group to estimate the specificity of the marker candidate (lung, breast, prostate, bladder, kidney, ovary, cervix, uterus, and other cancers without colonoscopy). In the final calculations of the specificity, these "control cancers" of subgroup II will be represented by appropriate mathematical models according to their prevalences given for the 50–59-year age group in the files for "Surveillance Epidemiology and End Results" (www.seer.cancer.gov) to mirror a screening situation. Subgroup III (100) contains gastrointestinal cancers (stomach, pancreas, esophagus, and liver). Subgroup IV (130) contains samples from patients with diverticulosis, diverticulitis, and inflammatory bowl disease including colitis with colonoscopy. Subgroup V (150) contains "healthy controls" with no bowel diseases, hemorrhoids, or bowel diseases other than CRC (note that these patients are not healthy *per se* because their complaints led to visits to have colonoscopy performed). Subgroup VI (150) contains samples from patients with adenoma grouped into adenomas $>$1 cm and $<$1 cm. Subgroup VII (200) contains benign specificity controls, 25 each, including autoimmune diseases, rheumatoid arthritis (RA), osteoarthritis (OA), hepatitis B, liver cirrhosis, chronic renal diseases, chronic obstructive pulmonary disease, and congestive heart failure. We deliberately omitted age-matched asymptomatic blood donors at this validation stage in panel B because we could not persuade enough donors to volunteer for a colonoscopy.

Panel B (based on its present composition) does not in any way mirror a screening situation. The cancer cohort, like the control cancers, is far overrepresented when compared with a true screening situation. For the sake of the work flow, this compromise must be accepted, or the test panel would have to be comprised of tens of thousands of samples. Even this

unmanageable number (within a validation screen) would not ensure a statistically sound number of CRC cases in the cohort, thus directly impacting the validity of the sensitivity estimates. Instead panel B can be viewed as a deliberately compiled "worst case" scenario panel. The immediate advantage of testing panel B in its present configuration lies in the fact that any shortcomings of the performance of a marker candidate become obvious *before* the marker is transferred to the clinical evaluation in large scale α-site studies. Only studies of this scale will reflect the true incidence and prevalence rates within a population and will be the ultimate measure of marker performance. We have established similar panel B collections for all indication areas in which we are conducting proteomic discovery programs to be able to adhere to the existing rigid validation schemes. We are expecting that the actual composition of these panels will eventually have to be modified once medical experts with in-depth medical knowledge in a particular field along with statisticians try to arrive at consensus compositions.

At any rate, given the number of proteomic candidate markers undergoing or awaiting validation, one can only call for standardized panels with a representative disease and control composition for any given disease under investigation. We anticipate that, in the absence of universal sample banks (which are not likely to be realized due to intellectual property questions associated with such universal sample banks), compulsory master panels will soon be introduced by committees within the proteomics community for all relevant disease areas. These master panels could be modeled according to panel B (as shown for CRC) and should be made available in the beginning for at least the major oncology areas (lung, colon, breast, and prostate), central nervous system diseases, and rheumatic and metabolic diseases. This would enable each investigator to collect the correct number of samples as specified in the relevant master list on their own and thus obtain a panel with approved specifications. Future claims regarding sensitivities and specificities should be based on those standardized sample collectives. Mandatory adherence to these basic requirements could well be the beginning of a process that would finally facilitate a comparison between marker performances reported by different groups with different test formats. Compliance should not be an issue because, at the end of an introductory period, the publication of future sensitivity and specificity claims would depend on the use of the master panel for the disease in question.

CONSIDERATION 7: REPORTING RESULTS

The data obtained in panel B above for each candidate is evaluated in our validation programs using receiver-operator characteristic (ROC) curves to calculate sensitivity at given specificities as outlined earlier (1). It must be repeatedly stressed that the composition of the test panels will dramatically influence the results. As a case in point, a given marker A identified in the rheumatoid arthritis program had a spectacular area under the curve (AUC) value of 0.97 when 389 RA patients were compared with 200 age-matched healthy blood donors. Adding some 200 OA patients to the control group resulted in the flattening of the ROC curve and a corresponding drop in the AUC to 0.78. Once the entire group of healthy donors was removed and the RA collective was directly compared with OA, the marker candidate resulted in an AUC of 0.58. This value is close to the diagonal in the ROC plots, which is equivalent to tossing a coin.[5] If one were to report only the results obtained in the healthy control group, one could correctly claim that marker A differentiates RA patients from healthy controls better than any marker currently in use. Healthy people without joint pain will rarely undergo testing for RA, so healthy controls in an RA evaluation panel are meaningless.

In the context of sensitivity and specificity claims one sometimes is at a loss to understand the absoluteness with which the data are presented. A short recollection of basic statistical facts might be helpful. Whatever sample number has been used to arrive at a sensitivity value (p), this value is only an approximation to the true value, which lies within a specified confidence interval (CI). The width of the CI scales with the S.E., *i.e.* the square root of the quotient of p times $(1 - p)$ divided by the sample size n. Using the normal approximation, the CI at a 95% confidence level equals $p \pm$ 1.96 times S.E. The direct consequence of these simple relationships is that when measuring 50, 100, 500, 1,000, or 10,000 samples, the resulting CI ($\pm$percentage of the measured sensitivity) will be (for $p = 0.8$) 11, 8, 4, 2, and 1%. If one measures only 50 samples (and some studies will in fact use this low sample number) and claims a sensitivity of 80%, then the true value lies (with 95% confidence) anywhere between 69 and 91%. If one takes the asymmetric distribution into account, the figures get worse (65–91%). Assuming the sensitivity is found to be only 60% based on the measurement of 50 samples, the CI (at 95%) will be $\pm$14%, or the true sensitivity values will be anywhere between 46 and 74%, a diagnostically irrelevant result. It would lend much needed credibility to reports based on activities in proteomics if it were made mandatory that any future marker performance claim inevitably be connected with the sample numbers, the sample types measured (composition of master panels!), and the resulting confidence intervals. It would greatly help to put claims into perspective and put an end to the public being misled by simplified headlines (down to the tabloid level) based on data sets excluding critical controls and not even meeting the most rudimentary statistical considerations.

CONSIDERATION 8: SUMMARY THOUGHTS

Altogether in my opinion, the track record of activities in proteomics with the diagnostic focus of successfully identifying biomarkers, validating them, and ultimately making them

[5] N. Wild, personal communication.

available to the medical community is modest to dismal. I believe that in the past there has been too much emphasis on the discovery phase representing only the first step of the entire process chain to arrive at diagnostically relevant biomarkers. Differences in the protein composition of two biological states are easily identified independently of the analytical technologies used. One dilemma of the present situation is traced to the fact that no convincing and generally accepted standardization regimen for the different technologies has been implemented. Compliance with such a standardized regimen could well be the basis for meaningful comparisons of different discovery approaches by different investigators. The simple PRS described might be one component of such a standardization system.

Few novel marker candidates have undergone rigorous and sufficiently detailed validation schemes tailored to the complexity of the disease in question. The simple notion that a marker candidate can in fact discriminate biological states in non-representative sample panels cannot and should not be the ultimate goal of the efforts of proteomics. Instead testing the marker candidate with a structured and rigorous validation concept must be accepted by the proteomics community in the future as the real and ultimate challenge. One much needed tool to put claims into perspective would be the mandatory use of consensus master sample panels as suggested.

The presentation in Asilomar that this contribution is in part based upon was originally entitled "The Proteomics Search for Diagnostic Biomarkers: Lost in Translation?". The views put forth here have been increasingly propagated by financial analysts and management teams in charge of selecting and supporting large scale and long term research and development programs in industry. In my opinion, it would be disastrous if the proteomics community did not counteract these ideas by introducing a more serious management of expectations based on *reproducibility* and *validation of results in relevant settings*. Without such a shift in emphasis, there is a real danger that both medical needs and commercial expectations will go unfulfilled.

Acknowledgments—I thank the members of the proteomics team in Penzberg for discussing and clarifying some of the issues raised. I am grateful to Stephan Palme, Norbert Wild, Michael Thierolf, and Gabriele Pestlin for providing unpublished data from ongoing research programs; Christoph Berding and Friedemann Krause (Biometry Department) for help in clarifying statistical questions; and Michael Tacke, Markus Roessler, Johann Karl, Marie-Luise Hagmann, and Gabriele Pestlin for critical comments and suggestions.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed. Tel.: 49-8856-60-4145; Fax: 49-8856-60-4513; E-mail: werner.zolg@roche.com.

REFERENCES

1. Zolg, J. W., and Langen, H. (2004) How industry is approaching the search for new diagnostic markers and biomarkers. *Mol. Cell. Proteomics* **3,** 345–354
2. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867
3. Pepe, M. S., Etzioni, R., Feng, Z., Potter, J. D., Thompson, M. L., Thornquist, M., Winget, M., and Yasui, Y. (2001) Phases of biomarker development for early detection of cancer. *J. Natl. Cancer Inst.* **93,** 1054–1061
4. Anderson, N. L. (2005) The role of multiple proteomics platforms in a pipeline for new diagnostics. *Mol. Cell. Proteomics* **4,** 1441–1444
5. Roessler, M., Rollinger, W., Palme, S., Hagmann, M. L., Berndt, P., Engel, A. M., Schneidinger, B., Pfeffer, M., Andres, A., Karl, J., Bodenmüller, H., Rüschoff, J., Henkel, T., Rohr, G., Rossol, S., Rösch, W., Langen, H., Zolg, W., and Tacke, M. (2005) Identification of nicotinamide *N*-methyltransferase as a novel serum tumor marker for colorectal cancer. *Clin. Cancer Res.* **11,** 6550–6557
6. Schmidt, A., Kellermann, J., and Lottspeich, F. (2005) A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* **5,** 4–15
7. Uhlén, M., Björling, E., Agaton, C., Al-Khalili Szigyarto, C., Amini, B., Andersen, E., Andersson, A., Angelidou, P., Asplund, A., Asplund, C., Berglund, L., Bergström, K., Brumer, H., Cerjan, D., Ekström, M., Elobeid, A., Eriksson, C., Fagerberg, L., Falk, R., Fall, J., Forsberg, M., Björklund, M. G., Gumbel, K., Halimi, A., Hallin, I., Hamsten, C., Hansson, M., Hedhammar, M., Hercules, G., Kampf, C., Larsson, K., Lindskog, M., Lodewyckx, W., Lund, J., Lundeberg, J., Magnusson, K., Malm, E., Nilsson, P., Ödling, J., Oksvold, P., Olsson, I., Öster, E., Ottosson, J., Paavilainen, L., Persson, A., Rimini, R., Rockberg, J., Runeson, M., Sivertsson, Å., Sköllermo, A., Steen, J., Stenvall, M., Sterky, F., Strömberg, S., Sundberg, M., Tegel, H., Tourle, S., Wahlund, E., Waldén, A., Wan, J., Wernérus, H., Westberg, J., Wester, K., Wrethagen, U., Xu, L. L., Hober, S., and Pontén, F. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* **4,** 1920–1932
8. Liao, H., Wu, J., Kuhn, E., Chin, W., Chang, B., Jones, M. D., O'Neil, S., Clauser, K. R., Karl J, Hasler, F., Roubenoff, R. Zolg, W., and Guild, B. C. (2004) Use of mass spectrometry to identify protein biomarkers of disease severity in the synovial fluid and serum of patients with rheumatoid arthritis. *Arthritis Rheum.* **50,** 3792–3803
9. Haab, B. B., Geierstanger, B. H., Michailidis, G., Vitzthum, F., Forrester, S., Okon, R., Saviranta, P., Brinker, A., Sorette, M., Perlee, L., Suresh, S., Drwal, G., Adkins, J. N., and Omenn, G. S. (2005) Immunoassay and antibody microarray analysis of the HUPO Plasma Proteome Project reference specimens: systematic variation between sample types and calibration of mass spectrometry data. *Proteomics* **5,** 3278–3291