Review

N. Leigh Anderson
Norman G. Anderson

Large Scale Biology Corporation,
Rockville, MD, USA

Proteome and proteomics: New technologies, new concepts, and new words

The goal of proteomics is a comprehensive, quantitative description of protein expression and its changes under the influence of biological perturbations such as disease or drug treatment. Quantitative analysis of protein expression data obtained by high-throughput methods has led us to define the concept of "regulatory homology" and use it to begin to elucidate the basic structure of gene expression control *in vivo*. Such investigations lay the groundwork for construction of comprehensive databases of mechanisms (cataloguing possible biological outcomes), the next logical step after the soon to be completed cataloguing of genes and gene products. Mechanism databases provide a roadmap towards effective therapeutic intervention that is more direct than that offered by conventional genomics approaches.

Contents

1 Introduction

The word "proteome", coined by Wilkins, *et al.* [1] in 1996, crystallizes an important concept whose counterpart for man was introduced more than 15 years ago [2–4], during the initial development phase of two-dimensional electrophoresis. The idea of a finite totality, comprising the functional (protein) molecular specifications of a genome, is a powerful token of the "solvability" (in principle) of the complex architecture of cells. As such, the word signals a renewed confidence among students of proteins, who have been somewhat displaced of late by the students of nucleic acids. The derived word "proteomics", which has come into use almost as an afterthought, may be equally useful since it indicates something less well-defined but more ambitious. Proteomics is a field, just as genomics is, rather than a closed and conceptually static body of knowledge (as are the genome and proteome, by definition). We define proteomics as: "the use of quantitative protein-level measurements of gene expression to characterize biological processes (*e.g.*, disease processes and drug effects) and decipher the mechanisms of gene expression control". As such, proteomics focuses on the dynamic description of gene regulation and, by doing so, offers something much more powerful than a protein equivalent of DNA databases: the concept of molecular regulation as a systematic science. For this reason, proteomics emphasizes quantitation and the assembly of large bodies of experimental observations in numerical databases. In this paper, we would like to briefly describe some of the history of what has now become "proteome and proteomics", discuss the relationships of these concepts to the DNA revolution, and indicate some of the most fruitful directions for future exploration.

Correspondence: Dr. N. Leigh Anderson, Large Scale Biology Corp., 9620 Medical Center Drive, Rockville, MD 20850-3338, USA (Tel: +301-424-5989; Fax: +301-762-4892; E-mail: leigh@lsbc.com)

Keywords: Proteomics / Genomics / Two-dimensional poylacrylamide gel electrophoresis

2 Protoproteomics

In 1975, when Klose's, O'Farrell's and Scheele's papers appeared describing high resolution two-dimensional electrophoretic methods [5–7], it seemed evident that systematic application of this approach would lead to the conquest of a new world of knowledge: the detailed workings of cellular machines. At that time, before the development of modern DNA methods (sequencing, cloning, recombinant techniques, and PCR), two-dimensional electrophoresis (2-DE) seemed to be the only tractable approach for surveying biologial complexity at the molecular level. Hence we set out, beginning in 1976, to develop technology [8–13], software [14, 15] and ideas that would allow a systematic enumeration of human proteins to support construction of the biological equivalent of the periodic table for man: the Human Protein Index (HPI) [2]. This effort led to the initiation of our "Molecular Anatomy Program" at the Argonne National Laboratory and the formation in 1980 of the "Human Protein Index Task Force", under the chairmanship of Norman G. Anderson, which prepared a report on the

0173-0835/98/1111-1853 $17.50+.50/0

project for the then-majority whip of the US Senate, Alan Cranston [16]. The objective of this effort, which included discussions with the National Institutes of Health, Department of Energy, National Aeronautics and Space Administration and major industrial companies, was the organization of a systematic scientific attack on the human proteome and the technology needed to resolve it. As it happened, major political changes took place that year (with the election of Ronald Reagan) that eliminated the political consensus required to drive the new initiative. The HPI project (or Human Proteome, as it would now be called), failed to attract large-scale support in the early 1980s, partly because large-scale science was then considered inappropriate in biology (the famous "fishing expedition") and partly because of failure to anticipate the feasibility and seductiveness of genomics. As Ivan Lefkovits has pointed out, if nature had not serendipitously provided us with restriction enzymes for the dissection of DNA, the "-omics" revolution probably would have been carried out with proteins first: proteome would have preceded genome and some later version of the HPI would have been undertaken. However, another scientific revolution intervened that propelled molecular biology in a different direction.

3 The genomics interregnum

DNA technologies intervened. They were powerful and easy to use. Progress was rapid and accelerating. The fashion in molecular biology turned from the generalist tastes of its founders to focus almost exclusively on nucleic acids. The DNA approach has something in it that resembles a mathematical formalism, and biology, always starved for an aspect that could be called legitimately theoretical, embraced this formalism as ground truth. Bioinformatics, the theoretical legacy of genomics, was built on the notion of string-shaped signals, pure information uncontaminated by messy chemistry and whose only real structure was a simple linear code deciphered at the dawn of the New Biology. Also, it could be pursued with a glorified word processor. Both user-friendly and chic: the "killer app" of biology. However, DNA is not the true bottom line: every modern textbook of biology explains that proteins embody the active life of cells, while nucleic acids represent only plans. There is more to paella than the recipe, more to Bach than ink on paper, and more to a society than its code of laws. In each case the implementation of a series of instructions is far more complex and interesting than one would expect from simply reading them. Thus it comes as no surprise that the pattern of mRNA abundances does not translate directly, or even approximately, into a description of the pattern of protein abundances on which cell behavior depends [17]. Cellular reality is more elaborate than the dreams of even the nucleus itself.

4 Why mRNA measurements cannot substitute for proteomics

To those of us who have worked in the protein domain for some time, it seems superfluous to mention the reasons why they must be included in an exploration of function. However, the indirect methods of so-called

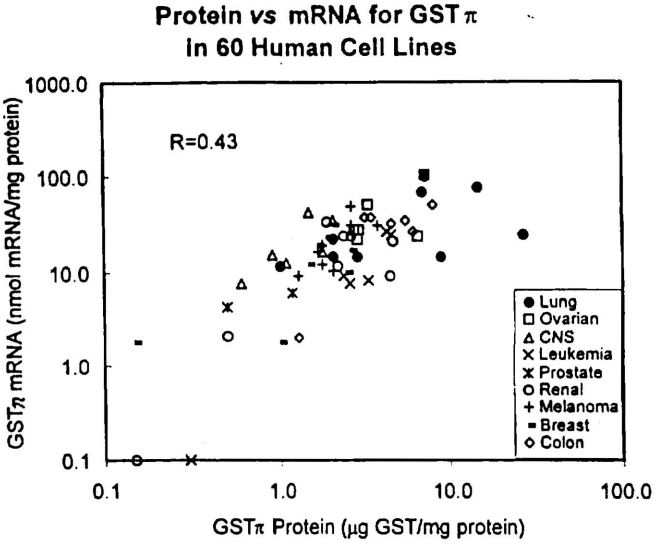

Figure 1. Data from Tew *et al.* [18] plotted to show protein abundance *versus* mRNA abundance for the π isozyme of glutathione-*S*-transferase in 60 human tumor cell lines on a log-log scale. A Pearson product-moment correlation was computed by the present authors, yielding a value of 0.43.

functional genomics have become so entrenched that we may be forced to reeducate a generation of molecular biologists in the "justification" for protein-level studies. Some of the principal arguments can be summarized as follows.

4.1 Protein-mRNA comparisons

At the last Siena meeting, N. Leigh Anderson, in collaboration with Jeff Seilhamer, presented the first multi-gene comparison plot of mRNA *vs.* protein abundances for cellular gene products, and found a correlation coefficient of 0.48 between them [17]. This result has occasioned considerable comment, ranging from consternation that it is so low, to amazement that it is so high. As in all quantitative science, the value obtained is, of course, subject to revision as more data is collected with better measurement methods. However, we believe that this result, halfway between a perfect correlation and no correlation at all, provides a reasonable benchmark. More recently we have plotted additional comparative data obtained by Tew *et al.* [18] (Fig. 1), comparing protein and mRNA abundances for one gene product across 60 human cell lines. The result here is a correlation coefficient of 0.43, and the plot shows that 10-fold variations in either protein or mRNA can be observed at constant values of the other parameter. In this study, protein was measured by an immunoaffinity-HPLC method and mRNA by quantitative Northern analysis – methods quite different from the 2-D gels and expressed sequence tag (EST) counting used in our initial comparison. Thus the result we obtained earlier seems not to be attributable to some simple methodological problem.

So far we thus have comparisons of mRNA to protein in one tissue across many genes, and for one gene across many cell types, both giving poor (0.5 or lower) correlations. These results refer to approximately static expression levels, however, and do not tell us whether changes

in protein and mRNA levels caused by a drug or a disease process will be better correlated. Experiments currently underway in several laboratories should answer this question definitively, through comparison of DNA array hybridization and 2-D gel data. Our expectation, based on numerous specific published instances comparing an mRNA and its protein in a single biological treatment system (*e.g.* [19, 20]) is that such correlations will also be low.

If this expectation is borne out, then the necessity to measure protein levels is inescapable. At the very least, mRNA increases or decreases in response to a biological perturbation must be systematically verified at the protein level. While many of the discrepancies between mRNA and protein levels may be the result of a derivative:integral relationship between the two, even this is an oversimplification since protein maturation and degradation are actively controlled as well [21]. Given the number of such effects, the question then becomes one of the relative levels of effort appropriate to find potentially misleading mRNA-level effects versus concentrating on proteins from the start.

4.2 Cellular control systems can operate purely in the protein domain without any mRNA involvement

A substantial fraction of interesting cellular regulation cannot be observed at the mRNA level by any technology, because the systems involved operate entirely in the protein domain. Recent work elucidating the mechanisms of 7-transmembrane G-protein coupled receptors (7-TM GPCRs) suggests that these systems are very numerous (thousands of receptors in man), and that they operate primarily through phosphorylation/dephosphorylation and migration of proteins [22]. Likewise, proteolytic modifications of membrane-bound precursors appear to regulate the release (functional expression) of a large series of extracellular signals such as angiotensin, tumor necrosis factor, various interleukins, and the Alzheimer's amyloid precursor protein [23]. This class of protein-level event is of great importance in pharmaceutical development, since it represents a potentially straightforward opportunity to affect a signaling pathway at an early critical step. Drugs acting by such mechanisms are, in fact, already in widespread use: angiotensin converting enzyme (ACE) inhibitors used in lowering blood pressure act by preventing the proteolytic conversion of a precursor to angiotensin [24].

4.3 Protein is more stable in many clinical samples than mRNA

Because they are disposable copies of genetic information, mRNAs are much more labile than DNA, both in terms of "spontaneous" chemical degradation (due to the possession of two adjacent hydroxyl groups on the ribose sugar ring) and in terms of the potency and ubiquity of degradative enzymes (the ubiquitous RNAse). Proteins are generally more stable, and exhibit generally slower turnover in most tissues. This disparity in stability has important consequences with regard to measurement of these molecules in biological materials, particularly clinical samples. Yolken and Johnston [25] have shown for example that mRNA levels can decrease almost 200-fold in human brain during a 48 h postmortem period, while with the same samples, we have shown that little, if any, decrease in total protein of native molecular weight is observed. Large losses of mRNA raise important questions with regard to the sequence-dependence of this effect, since it is probably unlikely that all lengths and sequences of message will decay at the same rate. Any nonuniformity of mRNA degradation thus introduces quantitative biases that grow larger and more complex as a function of time after the onset of tissue stress or death. While certain high-turnover protein modifications (*e.g.*, phosphorylations) and short half-life proteins can show postmortem changes, they are nevertheless likely to be more restricted than effects on mRNA.

4.4 Functions: proteins 100 000 – mRNA 1

The final, and perhaps most potent, argument in favor of protein measurements of gene expression has to do with function. Proteins implement almost all controlled biological functions, and hence are immediately involved in all important normal activities, disease processes, and drug effects. Messenger RNA is only that: a disposable message, having no other function than to serve temporarily to convey a piece of information from one place to another while being operated upon by proteins. mRNA measurements are therefore by definition indirect, while protein measurements relate directly to functional mechanisms. Although protein abundances are not quite as useful as direct measurements of all the cellular functions *per se*, we believe that such biochemical function measurements are not currently suitable for systematic measurement by any existing method, and thus that protein observations must suffice for the foreseeable future. From this perspective, the current term "functional genomics" (which implies that function can be explored at the genomic level), is a bizarre, probably oxymoronic, construction.

5 Regulation and function

Our proposed definition of proteomics is based on the notion that variations in the abundance and properties of proteins will allow us to observe what they are doing. For cells to operate properly, we have postulated [26] that essentially all genes are actively regulated, *i.e.*, that truly "constitutive" synthesis is the exception rather than the rule. Assuming evolution exerts significant pressure at the biochemical level, then gene regulation mechanisms ought to reflect functional relationships among genes. A variety of data support this view.

5.1 Protein expression effects reveal drug mechanisms

In an extensive series of *in vivo* studies of drug effects, we and others have observed that proteins whose abundance or structure is strongly regulated by a drug provide direct pointers towards a plausible mechanism of drug action. For example, dithiolethiones that protect

Table 1. A series of compounds whose effects in rat liver have been examined by the authors using proteomics methods to build a comparative drug effects database

Class	Generic name	Tradename	
1	5-Alpha-reductase inhibitor	Finasteride	Proscar
2	Angiotensin converting enzyme (ACE) inhibitor	Captopril	Capoten
3		Enalapril maleate	Vasotec
4	Acne product	Isotretinoin	Accutane
5	Adrenal steroid inhibitors	Aminoglutethimide	Cytadren
6	Alzheimer's treatment	Tacrine HCl	Cognex
7	Analgesic	Acetaminophen	Tylenol
8	Androgen	Stanozolol	Winstrol
9	Anesthetic	Halothane	Fluothane
10	Anti tuberculosis	Isoniazid	Nydrazid
11	Antibiotic	Tetracycline hydrochloride	Sumycin
12		Erythromycin estolate	Ilosone
13	Anticonvulsant	Valproic acid	Depakene
14	Antiestrogen, nonsteroidal	Tamoxifen	Nolvadex
15	Antifungal	Ketoconazole	Nizoral
16	Antineoplastic	Amethopterin (MTX)	Methotrexate
17		Amethopterin (MTX)	Rheumatrex
18	Antiviral	Zidovudine (AZT)	Retrovir
19		Acyclovir	Zovirax
20	Ca channel blocker	Amlodipine besylate	Norvasc
21		Isradipine	DynaCirc
22		Verapamil HCl	Calan SR
23	Carbonic anhydrase inhibitor	Methazolamide	Neptazane
24	Diuretic (K-sparing)	Spironolactone	Aldactone
25	Estrogens	Conjugated estrogens	Premarin Oral
26	Gall stone dissolution	Chenodeoxycholic acid	Chenix
27	Gout remedy	Allopurinol	Zyloprim
28	Immunosuppressant	Cyclosporine	Sandimmune
29		Tacrolimus (FK506)	Prograf
30		Azathioprine	Imuran
31	Lipid-lowering agent	Probucol	Lorelco
32		Gemfibrozil	Lopid
33		Lovastatin	Mevacor
34		Simvastatin	Zocor
35		Fluvastatin	Lescol
36		Niacin (nicotinic acid)	Nicolar
37		Pravastatin sodium	Pravachol
38	Nicotine delivery system	Nicotine (transdermal)	Nicoderm
39	Nonsteroidal antiinflammatory drug (NSAID)	Diclofenac	Voltaren
40		Oxaprozin	Daypro
41		Piroxicam	Feldene
42		Naproxen	Aleve
43	Psychoactive	Alprazolam	Xanax
44		Diazepam	Valium
45		Fluoxetine hydrochloride	Prozac
46		Triazolam	Halcion
47	Rheumatoid arthritis disease-modifier	Hydroxychloroquine	Plaquenil
48		Sulfasalazine	Azulfidine
49		Penicillamine	Cuprimine
50	Skeletal muscle relaxants	Dantrolene	Dantrium
51	Thyroid replacement	Levothyroxine sodium	Synthroid

against aflatoxin-induced liver cancer strongly induce a protein later found to be the aflatoxin B1 aldehyde reductase (responsible for detoxification of this carcinogen) [27]. Lovastatin, an inhibitor of 3-hydroxy-3-methylglutaryl (HMG) CoA reductase, and thus a cholesterol-lowering therapeutic) strongly induces a protein identified as HMG-CoA synthase (the preceding enzyme in the same pathway) [28]. Cyclosporin (a nephrotoxic immunosupressant) strongly decreases expression of a kidney protein identified as calbindin 28 kd, a calcium buffer whose loss apparently accounts for failure to excrete calcium and consequent accumulation of calcium deposits in the kidney [29]. Halothane administration causes covalent modification of specific liver proteins (trifluoroacetylation [30]), and these modified proteins produce a potentially fatal hyperimmunity to such modifications in rare individuals. The histamine H1 receptor antagonist methapyrilene (withdrawn from pharmaceutical use following discovery of its potent nongenotoxic hepatocarcinogenicity in rats) causes covalent modification of a series of mitochondrial proteins [31, 32], pointing to the action of a reactive drug metabolite in the mitochondria, where it may mutagenize mitochondrial DNA. Etomoxir, an irreversible inhibitor of carnitine palmitoyltransferase I, causes accumulation in liver of the adipocyte differentiation-related protein (ADRP), a protein thought to "clothe" the lipid droplets that accumulate as a result of the drug's blockage of lipid metabo-

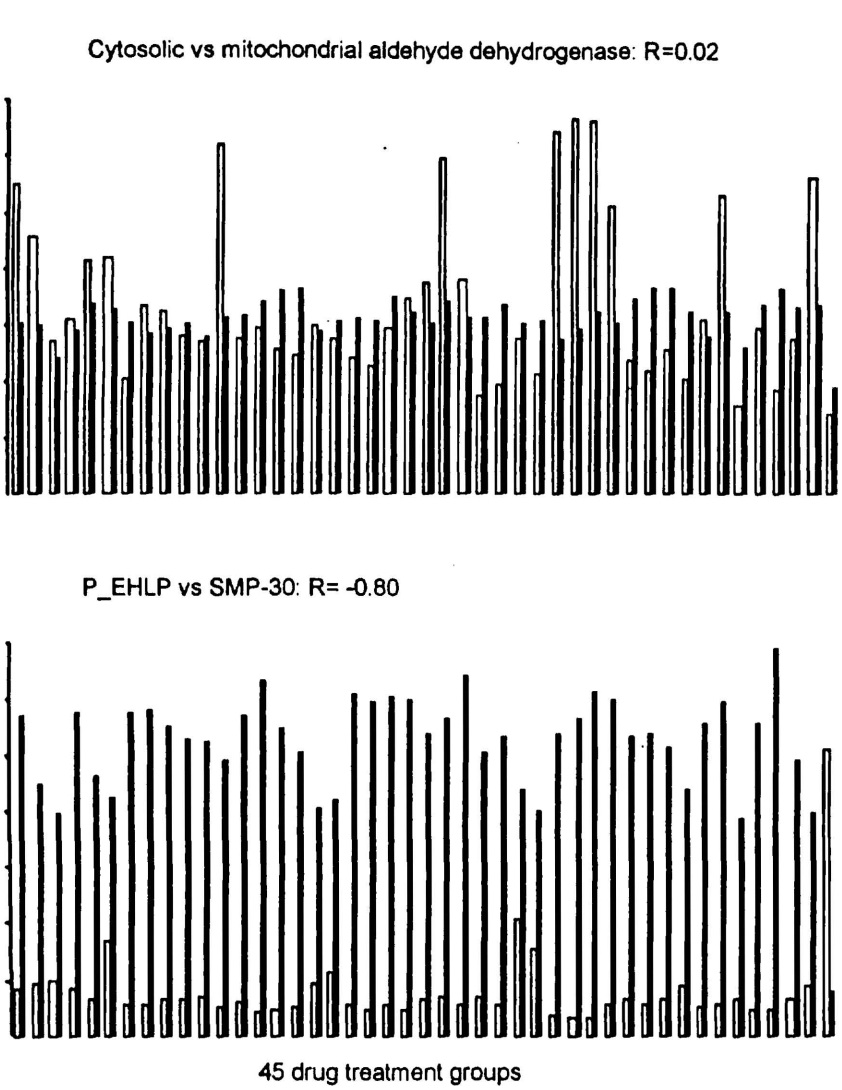

Figure 2. Bargraph comparisons of the abundances (plotted on an arbitrary vertical scale) of pairs of proteins in rat liver across 45 drug treatment groups.

lism [33]. These and other cases imply strong functional relationships between drug treatment, protein expression and resulting physiological effects.

Taken as a group, they suggest to us that protein-level effects may frequently constitute the drug or disease mechanism itself. This view is consistent with other indications that complex processes occur between drug administration and therapeutic effect. Psychoactive drugs, for example, often take several weeks to exert a therapeutic effect, despite the fact that a drug occupies the known receptor in the brain almost immediately. A slow, adaptive process of gene regulation subsequent to receptor binding offers a plausible explanation for this delay. Likewise, many drugs require an initial ramp-up to the desired dose, and slow de-escalation if therapy is to be discontinued, probably because the target cells modulate receptor density as the drug is introduced. The well-known properties of drug addiction and withdrawal, with an increasing threshold of drug effect, similarly suggest a gradual up-regulation of receptors and potentially other proteins. Cholesterol-lowering analogs of lovastatin, an inhibitor of the rate-limiting enzyme in cholesterol biosynthesis (HMG-CoA reductase), appear to lower blood cholesterol not so much because they decrease cholesterol synthesis, but because, as part of the hepatocyte's response to inhibition of the reductase, numerous other proteins are up-regulated, including the low density lipoproteins (LDL) receptor on the hepatocyte [34]. The latter effect may be primarily responsible for lowering blood cholesterol, and thus the therapeutic effect is a direct result of a drug-induced alteration in the abundance of a protein other than the drug target. In each of these cases, we find that the real therapeutic mechanism consists of the modulations in protein gene expression occurring as a secondary result of the initial binding action of the drug (usually to a protein identifiable as an enzyme, a receptor, or a membrane channel).

5.2 Regulatory homology *vs.* sequence homology in inferring function

This relationship between drug mechanisms and protein gene expression raises another important, and extremely useful, point. Drugs that act by similar mechanisms ought to produce similar protein gene expression effects. Different mechanisms should produce distinct effects. If this is true, as current evidence leads us to believe, then the pattern of protein changes should provide sufficient information to classify drugs according to their mechanisms of action (therapeutic or toxic), and allow a new and more sophisticated approach to the study of struc-

ture-activity relationships (SAR) – the backbone of medicinal chemistry. We are presently engaged in a systematic test of this view through the addition to our Molecular Effects of Drugs™ (MED™) database of *in vivo* effects of 51 pharmaceutical agents currently in medical use (Table 1). Preliminary analyses support the view that drugs of similar mechanism do indeed group together, based on protein effects.

The alternative approach, grouping proteins together on the basis of similarities in their regulation *via* drugs, also produces interesting results. Sets of proteins turn out to be coregulated, or anti-coregulated, under the influence of individual drugs (Fig. 2). These sets are sometimes surprising since the names of the proteins involved give no hint as to a reason for such relationships. P-EHLP (a peroxisomal enoyl hydratase-like protein) and SMP-30 (a cytosolic senescence marker protein), for example, show a correlation coefficient of −0.81 across 45 drug treatment groups in rat liver, which is a sign of strong inverse regulation. On the other hand, two proteins that might be imagined to show similar regulation do not: the cytosolic and mitochondrial aldehyde dehydrogenases show essentially no correlation (0.02) across these treatments.

The possibility of systematically obtaining regulatory correlations and anticorrelations between proteins provides us with a new category of homology that is potentially as important as sequence homology in tracing functional relationships. "Regulatory homology" (RH) measures the degree to which the cell attempts to coordinate the activities of two gene products, and hence is a good measure of whether the two function in some related way (*e.g.*, in the same pathway). Such control relationships are apt to experience the same selection pressures as protein sequences, and hence should evolve to reflect the optimal management of cellular performance. Instead of evolutionary lineage, however, RH traces functional optimization experience. As such, it provides a direct picture of what, from the cell's point of view, constitute linked functions. Since we know the functions of only a small fraction of the proteins, it should be possible to extend networks from the well-characterized to the uncharacterized proteins based on RH. The process thus implemented, which we call inference of function from regulation ("IFR"), can be carried out systematically on a large scale through the application of proteomics to a sufficient breadth of "regulating" situations.

5.3 Perturbation as a general approach to biological complexity

This approach can be generalized through consideration of the strategy one might use to deduce function inside the proverbial "black box" (Fig. 3). In physics, complex systems are frequently analyzed by what are called perturbation methods. A system, whose internal state we cannot adequately model *a priori*, is subjected to small perturbations in one or more input variables, and the effects on output variables are examined. The more input variables we can experiment with, the more completely we can model the system. Several ways can be envisioned for purposefully manipulating the input

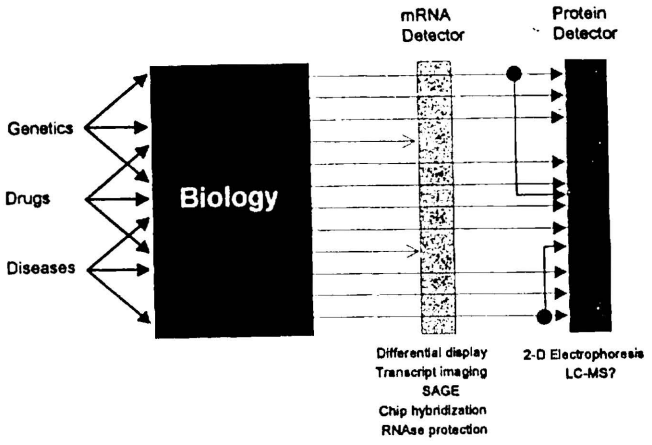

Figure 3. A schematic representation of the use of perturbation to investigate the internal structure of a biological system (a "black box"). Variable inputs produce variations in outputs which are observed by multichannel detectors at the mRNA or protein levels. Some variations in mRNA abundance may not yield differences in protein levels, while some additional protein alterations are caused by changes in other proteins (*e.g.*, by enzymatic post-translation modification).

variables used by the cell's control system. Classical genetics makes it clear that each gene can, in principle, be mutated or eliminated (saturation mapping), and the separate effects of these manipulations combined to build up a map of the relative "importance" of the gene. Such a project is now underway to characterize the genome of yeast through systematic knockouts. Likewise, the results with drugs make it clear that chemicals can be used for an equivalent purpose. The accepted basis of medicinal chemistry is that there should be at least one compound that specifically affects the function of each and every separate protein. Many if not most of these compounds may already exist in the combinatorial structure libraries used in pharma discovery, but are not recognized because most compounds are not screened for activity against more than a few protein targets. In any event, it appears that we have at least two comprehensive constellations of perturbations with which to explore the regulation of protein expression: genetic mutants and chemical structures. While the former is better established, the latter has the advantage of immediate application in the form of new lead compounds. A third group, environmental variables, provides a rich source of results, but without a systematic underlying structure so useful in designing large experiments.

6 Quantitative relationships between disease and therapy

Since many if not most therapeutic drugs act through mechanisms involving perturbations of protein expression, and since disease processes lead to protein changes as well, it is worth considering the relationship between disease and therapy at the protein expression level. In this context, we could define a perfect therapeutic as one that perfectly restores expression levels to the "normal" state (Fig. 4). Such perfect drugs are expected to be rare because the point of intervention (the target) is usually not the single point at which the disease defect occurs.

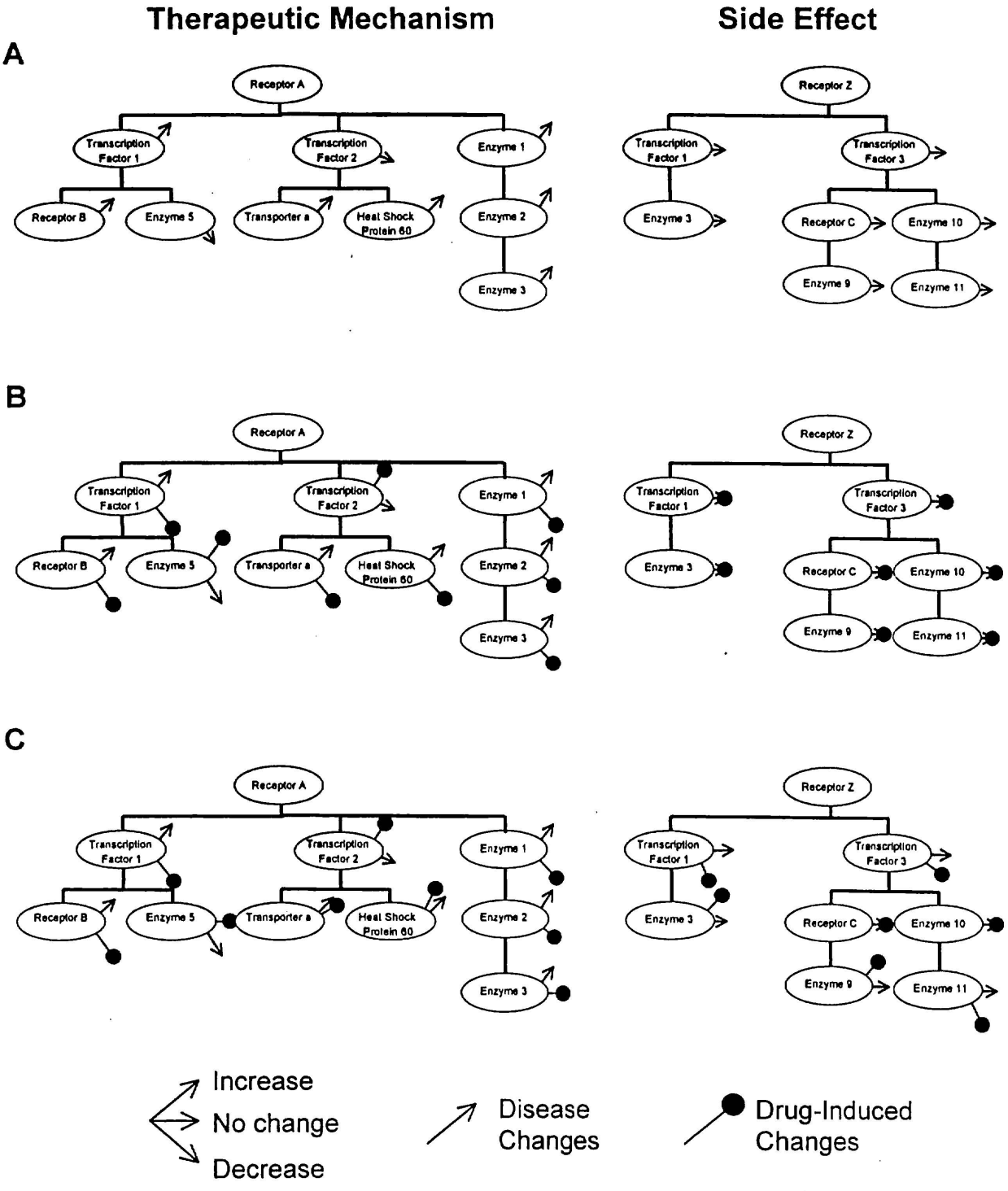

Figure 4. A schematic representation of effects on the expression levels of multiple proteins by (A) a disease, (B) a disease and a perfect drug, and (C) a disease and a real (imperfect) drug. Disease effects are shown as pointed arrows, and drug effects by arrows ending in balls. Increased protein abundance is symbolized by an upwardly directed arrow, and decreases by downward arrows, while no change is symbolized by a horizontal arrow. The disease causes changes in the pathway governed by receptor A (a pathway associated with the therapeutic mechanism), but not in a second pathway (governed by receptor Z) associated with a side effect. A perfect drug reverses the disease changes, and has no side effects (B). A typical drug (C) reverses some but not all of the disease effects, and induces undesired changes in a side-effect pathway.

Nevertheless, the concept has several useful applications. In the first place, drugs can be compared with respect to their effectiveness in restoring normal protein expression. For example, rheumatoid arthritis [35] and other acute phase inducers [36] cause quantitative abundance changes in a series of human serum proteins, thereby providing a series of potential indicators of disease status for use in trials of anti-inflammatory drugs. In a clinical trial comparison of piroxicam and tenidap, the latter drug proved to reverse more of the acute phase

changes than the former, providing an objective measure of "efficacy" at the molecular level [35]. While clinical trial success will always depend primarily on improvements in clinical signs, functional molecular measurements ("surrogate markers") provide quantifiable estimates of drug effects derived from detailed characterization of the disease process itself. This ability to quantitate provides potentially large rewards through improved comparison of drug efficacies.

Of potentially equal importance is the comparison of drug mechanisms. If all of a drug's effects are directed towards restoration of normal expression patterns, then the drug is likely to have fewer side effects. If, on the other hand, the drug also causes changes in proteins not affected by the disease state, changes that are too large (overcompensating for the disease) or in the wrong sense (amplifying the disease effect), then the drug is likely to be suboptimal. A series of candidate drugs could thus be ranked in terms of the overlap between their effects and the negative of the disease effect, thus producing a model of the SAR, which connects drug chemical structure and molecular effect. Likewise, drugs of equal therapeutic efficacy can be ranked by the number of proteins whose expression they affect – in this case, the fewer the better, on the assumption that a good drug causes the least perturbation consistent with adequate therapeutic effect.

7 Future challenges in proteomics

The principal technical challenge is to achieve a level of comprehensiveness in relation to proteins that corresponds to the situation in genomics: complete coverage. This is clearly a much harder job with proteins than with nucleic acids. In the first place, genes are approximately equimolar in genomic DNA, whereas proteins may span 7 or 8 orders of magnitude in functional abundance in a cell type such as the hepatocyte, and potentially wider ranges in distributed media such as serum. Very few detection methods exist that are usable over seven orders, and if one could be found, there remains the problem of the size of the Gaussian "tails" of abundant spots obscuring minor ones. In addition, we have the difficulty of resolving very hydrophobic, very basic, or very large proteins in current 2-D systems. The relative chemical homogeneity of DNA, the existence of reverse transcriptases, restriction enzymes, PCR and sequence complementarity have all contributed to the conspicuous ease of genomics as compared to proteomics. There may be a temptation to assert that complete proteomes are just around the corner – an assertion that represents a major leap of technological faith at this point since it has not so far been achieved even for a simple prokaryote. In our view, concern over the completion problem, while valid for proteome studies, is misplaced from the viewpoint of proteomics. In proteomics, major discoveries will be made through quantitative observations of a limited (but large) number of protein gene products once the database is rich enough.

In response to the technical challenges, we are likely to see the emergence of fully automated 2-D systems capable of tens of thousands of gels per year at very high resolution (one is under development at LSB). Ultimately (in perhaps 3–10 years), we will also see nongel-based alternative technologies, possibly using combinations of capillary electrophoresis or liquid chromatography with mass spectrometry, that may make proteomics data acquistion even more routine. And in the near future we will see systematic MS methods that allow identification of every spot on every gel, thereby completing the linkage to genomics, and simultaneously freeing us from dependence on any particular protein separation system.

The major intellectual challenge in proteomics is, of course, data analysis. What began as an attempt to produce annotated maps of proteins has evolved into a systematic effort to mine knowledge from broad measures of biological system performance. A wide variety of sophisticated approaches have been applied to proteomics data sets, including numerical taxonomy and multivariate statistics [37, 38], quantitative trait locus (QTL) mapping [39], heuristic clustering [40], similarity clustering [41] and regulatory homology (described here). The major limiting factor for all of these approaches has so far been the limited size of data sets available for analysis, and as this rapidly improves we are likely to see major advances in the results of such data mining. Of particular interest to us is the development of the database of gene expression control mechanisms: in terms of practical utility, this database is likely to prove more fruitful than the database of all the genes.

8 Conclusion

By the turn of the millennium, if not much sooner, we will see a dramatic shift of emphasis from DNA sequencing and mRNA profiling to proteomics. Considered objectively, there is every reason to expect that proteomics will ultimately exceed genomics in total effort, though this growth will be sorely limited by the availability of scientists able to deal with proteins' nonideal properties, with quantitative rather that qualitative (*e.g.*, Northern blot) data, and with the complex modeling formalisms that will predominate in discussions of gene regulation in the next century.

Received February 24, 1998

9 References

[1] Wilkins, M. R., Sanchez, J.-C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., Williams, K. L., *Biotechnol. Genet. Eng. Rev.* 1996, *13*, 19–50.
[2] Anderson, N. G., Anderson, L., *Clin. Chem.* 1982, *28*, 739–748.
[3] Anderson, N. L., *Trends Anal. Chem.* 1982, *1*, 131–135.
[4] Taylor, J., Anderson, N. L., Scandora Jr., A. E., Willard, K. E., Anderson, N. G., *Clin. Chem.* 1982, *28*, 861–866.
[5] Klose, J., *Humangenetik* 1975, *26*, 231–243.
[6] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007–4021.
[7] Scheele, G. A., *J. Biol. Chem.* 1975, *250*, 5375–5385.
[8] Anderson, N. G., Anderson, L., *Anal. Biochem.* 1978, *85*, 341–354.
[9] Anderson, N. G., Anderson, L., *Anal. Biochem.* 1978, *85*, 331–340.
[10] Lefkovits, I., Young, P., Kuhn, L., Kettman, J., Gemmell, A., Tollaksen, S., Anderson, L., Anderson, N. G., in: Lefkovits, I., Pernis, B. (Eds.), *Immunological Methods*, Academic Press, Orlando 1985, *Vol. 3*, pp. 163–185.

[11] Anderson, L., *Two-Dimensional Electrophoresis: Operation of the ISO-DALT® System*, Large Scale Biology Press, Washington, DC 1991.
[12] Pearson, T., Anderson, L., *Anal. Biochem.* 1980, *101*, 377–386.
[13] Anderson, N. L., Hickman, B. J., *Anal. Biochem.* 1979, *93*, 312–320.
[14] Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P., Anderson, N. G., *Clin. Chem.* 1981, *27*, 1807–1820.
[15] Anderson, L., in: Schafer-Nielsen, C. (Ed.), *Electrophoresis '88*, VCH Verlagsgesellschaft, Weinheim 1988, pp. 313–321.
[16] Anderson, N. G., Abajian, V., Anderson, N. L., Johnson, I., McConkey, E., McDermott, W., Neel. J. V., Thomas, S., Whitehead, E. C., *Report of the Human Protein Index Task Force*, Privately published, 1980.
[17] Anderson, L., Seilhamer, J., *Electrophoresis* 1997, *18*, 533–537.
[18] Tew, K. D., Monks, A., Barone, L., Rosser, D., Akerman, G., Montali, J. A., Wheatley, J. B., Schmidt Jr., D. E., *Mol. Pharmacol.* 1996, *50*, 149–159.
[19] Hoogeveen, R. C., Reaves, S. K., Lei, K. Y., *J. Nutr.* 1995, *125*, 2935–2944.
[20] Skidmore, A. F., Beebee, T. J., *FEBS Lett.* 1990, *270*, 67–70.
[21] Hargrove, J. L., Schmidt, F. H., *FASEB J.* 1989, *3*, 2360–2370.
[22] Wess, J., *FASEB J.* 1997, *11*, 346–354.
[23] Hooper, N. M., Karran, E. H., Turner, A. J., *Biochem. J.* 1997, *321*, 265–279.
[24] Pfeffer, M. A., *Am. Heart J.* 1993, *126*, 789–793.
[25] Yolken, R., Johnston, N., *J. Neurosci. Methods* 1997, *77*, 83–92.
[26] Anderson, N. G., Anderson, N. L., *Electrophoresis* 1996, *17*, 443–453.
[27] Anderson, L., Steele, V. K., Kelloff, G. J., Sharma, S., *J. Cell. Biochem.* 1995, *22*, 108–116.
[28] Anderson, N. L., Esquer-Blaso, R., Hofmann, J.-P., Anderson, N. G., *Electrophoresis* 1991, *12*, 907–930.
[29] Steiner, S., Aicher, L., Cordier, A., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., *Biochem. Pharmacol.* 1996, *51*, 253–258.
[30] Kenna, J. G., Satoh, H., Christ, D. D., Pohl, L. R., *J.Pharmacol. Exp. Ther.* 1988, *245*, 1103–1109.
[31] Anderson, N. L., Copple, D. C., Bendele, R. A., Probst, G. S., Richardson, F. C., *Fundam. Appl. Toxicol.* 1992, *18*, 570–580.
[32] Cunningham, M. L., Pippin, L. L., Anderson, N. L., Wenk, M. L., *Toxicol. Appl. Pharmacol.* 1995, *131*, 216–223.
[33] Steiner, S., Wahl, D., Mangold, B., Robison, R., Raymackers, J., Meheus, L., Anderson, L., Cordier, A., *Biochem. Biophys. Res. Comm.* 1996, *218*, 777–782.
[34] Bilheimer, D. W., Grundy, S. M., Brown, M. S., Goldstein, J. L., *Proc. Nat. Acad. Sci. USA* 1983, *80*, 4124–4128.
[35] Doherty, N. S., Littman, B. H., Reilly, K., Swindell, A. C., Buss, J. M., Anderson, N. L., *Electrophoresis* 1998, *19*, 355–363.
[36] Bini, L., Magi, B., Cellesi, C., Rossolini, A., Pallini, V., *Electrophoresis* 1992, *13*, 743–746.
[37] Anderson, N. L., Hofmann, J.-P., Gemmell, A., Taylor, J., *Clin. Chem.* 1984, *30*, 2031–2036.
[38] Tarroux, P., *Electrophoresis* 1983, *4*, 63–70.
[39] Damerval, C., Maurice, A., Josse, J. M., de Vienne, D., *Genetics* 1994, *137*, 289–301.
[40] Appel, R., Hochstrasser, D. F., Roch, C., Funk, M., Muller, A. F., Pellegrini, C., *Electrophoresis* 1988, *9*, 136–142.
[41] Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace Jr., A. J., Kohn, K. W., Fojo, T., Bates, S. E., Rubinstein, L. V., Anderson, N. L., Buolamwini, J. K., van Osdol, W. W., Monks, A. P., Scudiero, D. A., Sausville, E. A., Zaharevitz, D. W., Bunow, B., Viswanadhan, V. N., Johnson, G. S., Wittes, R. E., Paull, K. D., *Science* 1997, *275*, 343–349.