Review

George L. Gabor Miklos[1]
Ryszard Maleszka[2]

[1]GenetixXpress Sydney, Australia
[2]Visual Sciences, Research School of Biological Sciences,
The Australian National University,
Canberra, Australia

Integrating molecular medicine with functional proteomics: Realities and expectations

We analyze key proteomic issues and cutting-edge technologies that will spearhead inroads into functional interpretations of human diseases and their therapeutic rectification, following the availability of the predicted human proteome. We contrast the distinctions between high quality data that are low throughput, (*e.g.*, 3-D proteomic reconstructions in embryogenic and nervous system contexts, and multigenerational transgenic studies), *versus* automated data harvesting that is more distant from human disease phenotypes and currently fulfills a diagnostic role, (*e.g.*, molecular portraits of human diseases *via* transcriptomic analyses). We examine the extent to which these approaches impinge upon a realistic understanding of human diseases, namely how close they come to revealing the causal events involved in the initiation of disease. While tissue sources from human embryogenesis, foetal development and the brain remain the absolute priority, the pragmatic approaches utilize judicious data integration from selected proteomic studies of model organisms. The role of genome-wide disease-related screens, "humanized" transgenic analyses, multigenerational gene interference methods, and analyses of post-translational modifications in epigenetic contexts from *Drosophila* will be crucial, since these avenues are far too slow and transgenically cumbersome in mammals. Finally, the implementation of multi compartment electrolyzers (MCE) and multi photon detection (MPD) systems will be pivotal for the proteomic profiling of human tissue samples.

Keywords: Proteomic profiling / Human diseases / Review PRO 0010

Contents

Correspondence: Dr. George L. Gabor Miklos, GenetixXpress Proprietary Limited, 78 Pacific Road, Palm Beach, Sydney, NSW, AUSTRALIA, 2108
E-mail: genetixxpress@ozemail.com.au
Fax: +61-2-9974-3111

Abbreviations: SNP, single nucleotide polymorphism; **MCE,** multicompartment electrolyzer; **MPD,** multiphoton detection

1 Introduction

What priorities emerge when the predicted human proteome becomes available?

The completion of the sequencing phase of the human and mouse genomes [1] provides the opportunity for describing human diseases in terms of complete network perturbations instead of individual genes and proteins. What types of data are required to dissect these multilayered networks in the normal and diseased states and to move from molecular diagnostics to generating therapeutic targets? Advanced information acquisition is from three parallel sources; proteome projects, detailed structural network projects and integrated phenomics projects.

The first of these, the Human Proteome Project, aims to couple high-throughput technologies to annotation-driven comparative proteomics and is complete when all

1615-9853/01/0101–0030 $17.50+ .50/0

the post-translationally unmodified variant proteins arising from every transcription unit in the human genome are known.

The detailed structural network projects involve descriptions of the proteomic profiles from normal individuals (where such a profile comprises the population of all *in vivo* protein variants and their relative frequencies and post-translational modifications, as well as all variants that arise from intein-exein protein splicing and RNA editing), in every human cell type from fertilization onwards. This includes the biochemical and organismal evaluation of 3-D structures, their involvements in multisubunit complexes in all cell types, their intra- and extracellular locations, the departure points and destinations of their proteolytic cleavage products, and the results of their *in vivo* manipulation in specified epigenetic contexts over the normal life cycle.

The integrated phenomics phase is proteomic profiling of human diseases instantiated in cell biological, anatomical, physiological and phenotypic contexts. It is a description of network properties, reorganizations and their terminal equilibria from conception to death in normal *versus* diseased states and in age related phenomena. This phase captures the highest quality data and is only complete when molecular network knowledge is sufficiently robust to permit clinically precise predictions for the personalized therapeutic amelioration of a given disease.

Except for already available drug targets, it is not yet known which components of proteomic profiles are biologically relevant for human disease networks in different individuals or which are excellent therapeutic targets for a given disease. Hence the first task is a diagnostic one; to obtain the proteomic profiles of normal and diseased tissues and to biologically ascertain which protein combinations are the key contributors to these two categories in the specific genetic background of an individual.

2 Approaches to uncovering disease components

The molecular approach to disease analysis currently relies upon high-throughput microarray-driven differential display transcriptomics of diseases, and the analyses of single nucleotide polymorphisms (SNPs) in different human populations [2–5]. The hopes are that some transcriptomic data are sufficiently transferable to the protein level that adequate information will emerge to clarify their importance in a particular disease and allow for drug target prediction, and that the SNP data will highlight major disease susceptibility *loci*.

The proteomic approach involves gel-based, chip-based, capillary-based and mass spectroscopy-based differential display to uncover proteins whose levels and post-translational modifications differ between diseased and normal tissues [6, 7]. The hopes here are that these technologies can be ramped up in clinical settings to achieve sufficient throughput to move into the functional aspects of post-translational modifications, multisubunit complexes, intracellular compartmentalization, cell-cell interactions and the action of proteolytic cleavage products acting at great distances from their source of production elsewhere in the body.

The third approach to disease analysis and therapeutics is to use combinatorial chemistry to rapidly identify small molecules such as RNA aptamers and nonpeptides that bind to cellular components such as specific subtypes of receptors, and utilize these to ameliorate specifically chosen diseases [8, 9]. In essence, the goal is to have a companion small molecule for every protein product.

3 Molecular portraits of human diseases

The generation of molecular portraits of human diseases *via* differential display methods is critical for the patient, and provides finer subcategories than the pure histological or serum marker approaches that are currently in use for various forms of cancer. However, current interpretations of differential display data are often compromised by three optimistic rationalizations; the first is that those genes whose product levels fluctuate significantly between the normal and disease state are the ones causally involved in the disease under investigation; the second is that extrapolations from the transcriptome to the proteome are functionally robust; and the third is that differentially expressed genes and proteins are prime candidates for pharmaceutical targets. These are three high risk assumptions. For example, it is already clear from the first in-depth microarray analysis of prostate tumors that there is tremendous variation between individuals with prostatic adenocarcinomas. Hundreds of genes are over- and under-expressed in tumors when single individuals are examined, but the commonalities reduce to only a few dozen when more than ten individuals are analyzed [10]. Thus while prognostic markers are available for many cancers and are statistically significant for a group, they do not automatically provide the knowledgebase for the individual patient; that is, whether to proceed with radical surgery [11].

In addition, there is significant variation between biological networks that predispose to the same disease phenotype. While the possession of the apolipoprotein epsilon 4 allele is considered to be a major genetic risk factor for

Alzheimer's disease, it is now clear that this generalization simply does not hold up in certain populations [12]. Furthermore, although the deposition of beta amyloid is implicated in the pathophysiological cascade that predisposes to progressive neural shutdown, the between-individual variation in the number of cortical plaques is large [13]. While such deposits can occupy 15 % of the surface of the cortex, there is still a poor correlation between the clinical manifestations of the late onset form of the disease and the amount of cortical deposition.

One thing is certain, when crunch time comes around, decisions on the causal factors underpinning a given disease inevitably face only one question, the answer to which determines the years of research that will be invested in particular proteins;

– which of the proteins, (whose expression levels, cleavage products, post-translational modifications, or time and place of production and destination are altered in development and disease), are causally implicated in the processes under examination, and which are innocent bystanders that are peripheral to the disease, but whose alterations have been imposed upon them by network changes elsewhere in the system, or during the development of the disease?

Once a cohort of candidates is accepted, the amount of effort that goes into their detailed characterization, including knockout analyses in model organisms, 3-D structural analyses, post-translational modifications, anatomical localizations, and evaluations in the therapeutic pipeline, is truly immense. If the initial selection is flawed, the consequences are catastrophic; the results are oblique to the disease under examination, and research and pharmaceutical pipelines become clogged with irrelevant entities.

Therefore the pressing challenges in functional proteomics are to obtain human material relevant to the disease under investigation, and to evaluate the biological significance of candidates which emerge from differential display, from time series, from drug perturbations, and disease screens. In what follows, we examine some strategies that are helpful in integrating proteomics with molecular medicine in the context of phenomic and technological issues, and present a realistic assessment of what is required in achieving closure in understanding some human diseases. We start with the most challenging issue; what priorities emerge after the annotation of the predicted human proteome?

4 Obtaining relevant human material

A rough draft of the human proteome is already available from a combination of annotation-driven bioinformatics, comparative genomics and proteomic analyses. The next phase requires the generation of proteomic profiles from each of the hundreds of human cell types from individual human beings. This requires a serious reappraisal of how to sample early human development. Furthermore, since there are hundreds of morphologically recognizable human cell types, the challenges are to obtain clinical access to appropriate amounts of each cell type during the development and aging of normal individuals, and to generate the proteomic profile for each cell type of an individual.

The dynamic range of protein abundance spans at least six orders of magnitude, yet the critical issue for disease phenotype is the biological efficacy range necessary to maintain network stability. Since 90 % or so of the protein mass in a given cell type comes from only 10 % of the proteins, super sensitive technologies are required to even detect those proteins whose levels are very low, but whose biological potency is high, such as G-protein coupled receptors, transcription factors and kinases. Many well studied proteins in humans, mice and flies can be reduced to a few percent of their wild type levels without debilitating phenotypic consequences, and the amount of an individual protein can vary significantly between different wild type individuals. The extent to which this natural between-individual variation contributes to phenotype is not well documented.

In the context of the development and aging of the human nervous system, it is not yet clear how to best access its cell type complexity. The subdivisions of the neural real estate are neither easily identifiable in embryogenesis, nor in adult brains, and the interdigitation of neuronal cell types between different cortical areas and deeper brain regions is extensive. For example, the functioning of the multiple subtypes of the human serotonin receptor family underpins aspects of memory, depression, appetite control, thermoregulation, cardiovascular function, and nausea, yet to proteomically access this circuitry and its components is not facile. The two morphologically distinct serotonergic fiber systems that stream from the dorsal and median raphe nuclei to the cerebral cortex, the diencephalon, and telencephalon, are so pervasive, that almost every brain cell is close to a serotonergic fiber. Thus obtaining specific brain tissue from the serotonergic system of an embryo or foetus is experimentally daunting. Even in adults, accessing structures such as hypothalmi and pituitaries is a *post mortem* endeavor. While one alternative is to use mice and monkeys, significant differences sometimes exist between them and human beings. Mice do not suffer from Alzheimer's disease, and intriguingly, there are large spindle shaped neurons in the anterior cingulate cortex of humans and great apes that are not present in rodents. These unique neurons are particularly susceptible to degeneration in Alzheimer's patients [14].

Obtaining homogeneous material from patients presenting with various conditions also needs to be put into a clinical perspective. It is generally straightforward to obtain samples from leukemias, lymphomas, reduction mammoplasties and cutaneous melanomas, but impractical to obtain neural tissue from dorsal root ganglia, spinal cord or brain from patients suffering from chronic pain, anxiety disorders, severe migraine, or chronic neuropsychiatric disorders, such as Tourette's, autism, and schizophrenia. The brain material from these latter cases is invariably from *post mortem* brain samples, and the individuals have usually been treated with a variety of medications; a significant complication for the interpretation of proteomic profiling. In addition, for many conditions such as type II diabetes and hypertension, it is not clear what tissue needs to be sampled. It may be that in some cases multiple tissues need to be assayed (for diabetes; pancreas, liver and adipose tissue?), but what these multiple tissues may be for specific diseases is still unclear.

However, it is not just the brain that has a diversity of cell types; biopsy material from tumors generally consists of a cocktail of different cell types [11]. Ninety percent of adult human cancers are solid and contain blood vessels, inflammatory cells, necrotic tissue, stroma, and noncancerous cells, each of which will have its own characteristic proteomic profile. Cancers of the prostate routinely manifest as distinct multiple *foci*, not all of which are invasive, and hence they are molecularly variable. Proteomic analysis of different tumors from the same person will therefore require extremely sensitive detection technologies, as input material will always be limiting.

How can the proteomic profiling data be placed into relevant anatomical contexts?

5 The proteomics-cell biology interface

The best characterized cases of protein localizations and functions stem from the distribution of neurotransmitter and receptor distributions in mammals, where contrary to popular belief, transmitter and receptor localizations are more often out of register, than in concert [15]. For example, the caudate nucleus has no visible adrenergic fibers, yet this nucleus has the highest level of beta adrenergic receptors of the entire brain. The basolateral nucleus of the amygdala has a dense beta adrenergic receptor population but no norepinephrine containing fibers; by contrast, the central nucleus has a very dense population of fibers but very few receptors. The *substantia nigra pars reticulata* has extremely high levels of tachykinin-containing terminal immunoreactivity, but tachykinin receptors are virtually undetectable. Similarly, neurotensin receptors are essentially missing from the central nucleus of the amygdala, which nevertheless has the densest number of neurotensin containing fiber terminations. There is, furthermore, great variation between species in the distribution of receptor types. Cholecystokinin receptors are predominant in the cerebellum of the guinea pig, whereas cholecystokinin receptors are not detected in the cerebellum of the rat.

There is also quite often a serious misprision about what processes a protein ought to function in, in which tissues it ought to perform those functions, and that the pathology of a disease ought to reflect the sites of major gene expression. In the case of the Huntingtin gene, the transcript is heavily expressed in the brain mainly in the dentate gyrus and pyramidal neurons of the hippocampus, the cerebellar granule cell layer, the pontine nuclei, and in cerebellar Purkinje cells; however, the neuropathology of the disease is most prominent in the basal ganglia. Thus there is no simple correspondence between prominent RNA expression and neuropathology [16]. In other well known cases such as familial amyotrophic lateral sclerosis, there is extensive expression of *SOD*1 throughout the body, but the pathology is restricted to motor neurons in the motor cortex and spinal cord. In the case of Tenascin-C, with its prominent expression in ligaments and tendons, the mouse knockout experiment reveals that there are problems in serotonin and dopamine transmission in the cortex, hippocampus and striatum, but not with tendons and ligaments [17, 18].

In the case of the oncogenic c-src tyrosine kinase, which makes up a large component (0.2–0.4 %) of the total protein in platelets and in the growth cones of neurons, the major expectation was that it played a pivotal role in blood and brain. In fact, the blood and brain systems are normal in c-src knockout mice, but the mice suffer from osteopetrosis. In fact, c-src is expressed in osteoclasts and its expression there had gone unnoticed. It is likely that c-src is expressed in platelets and neurons owing to the default characteristics of the regulatory networks in which it is embedded. As long as it is not toxic or deleterious to platelets and neurons, its presence is tolerated [19].

These examples illustrate important principles which are largely absent from conventional interpretations; namely that most metazoan genes belong to overlapping sets of networks within each of which many of their protein products are not functionally required for the basic outputs of that network. Hence when one particular gene is absolutely needed in a tissue, the transcriptional machinery is likely to turn on many others even when their presence is not required. This is a default, or unavoidable consequence of the evolutionary history, and current structure of networks. Thus when it comes to microarray and proteomic data few investigators have doubts about the significance of the measurements, but some doubt their meaning for the biology of the cell [20].

Finally, the previous examples reveal that the critical knowledge acquisition concerns protein expression patterns and their functional evaluation in terms of phenotype. In this arena the hard yards have be gained by sheer biochemical and neuroanatomical grind; there are no short cuts, no quick-fix bioinformatic solutions, and no guarantees that the highest RNA and protein expression levels are indicators of potential sources of pathology. Nevertheless, the payoff from such low-throughput analyses is extraordinary; they are of extremely high functional quality.

6 Differential displays and disease candidates

The hyperbole usually associated with gene or protein candidates that have been freshly plucked from differential display experiments and thrust into the limelight as potential therapeutic candidates, leads many to believe that yet another important human disease has been solved. The overwhelming majority of these candidates quietly slip into obscurity as the biological blowtorch of function is slowly turned up. Why is this the case? The answer is quite straightforward. It is already known from transgenic data on knockout, knockin, misexpression, aptamer and antisense perturbations from *Caenorhabditis, Drosophila, Mus* and from limited human mutational data sources that at least 50 % of the major disturbances at the genomic level produce such small effects on phenotype that they are very difficult to measure under laboratory conditions [21–24]. It is unavoidable therefore that a significant proportion of the candidates that are harvested from any differential display experiment will fall into this category of phenotypic blandness. For example, when vimentin, a conspicuous and supposedly critical member of the cytoplasmic intermediate filament network of mammals is eliminated from all cells, the knockout mice are reproductively normal and show no obvious phenotypic effects [25]; the usually conspicuous vimentin filament network is totally missing and has not been compensated for by any other filament system. This example and a plethora of others reveal that many proteins are not of critical importance in a global context but they make small contributions to the efficiency and reliability of networks [26].

Thus no matter how important a protein may appear *a priori* to an investigator, and no matter how inventive we are in rationalizing its alterations in diseased tissue, or in associating it *via* cluster analysis with other familiar disease candidates only experimental perturbation reveals whether the changing proteomic profile is of fundamental importance to the genesis of a disease, or whether the fluctuations are a biologically insignificant default outcome of networks passing through unstable states before reaching equilibrium at stable attractors.

7 A realistic view of human diseases

Human “monogenic” diseases are such gross oversimplifications that they are predictively unhelpful. Even the classical textbook example of sickle cell anemia has revealed itself to be far more complex. Individuals harboring identical mutant alleles at the beta globin *locus* can have very different phenotypes, ranging from an unrecognizable departure from normalcy, to childhood mortality. This diversity in disease phenotype is seen, without exception, in every human disease that has received adequate investigation (Alzheimer, Hirschsprung, neurofibromatosis type 1, hemochromatosis, schizophrenia, psoriasis, cancer, autism, autosomal dominant optic atrophy, and so forth). One reason for this complexity in phenotypic variation is the enormous underlying variation at the anatomical and molecular levels that occurs between individuals, and the second reason is the degeneracy inherent in any network.

In humans and macaque monkeys, for example, variation in the size of the cortex and various architectonic areas is substantial; variation in single architectonic areas of the striate and extra striate cortex is of the order of 20–40 %. Even in monozygotic twins, the surface areas of the dorsolateral frontal cortex can vary significantly. In different mouse strains, the variation in neuron number in the *locus coeruleus*, *substantia nigra*, cerebellum, olfactory bulb, hippocampus and neocortex can be of the order of 25–100 % [27]. The consequences of this between-individual variation are that measurements at the proteomic level can be quite misleading, and need to be complemented with neuroanatomical data in order to avoid misinterpretations of the differential display data. For example, measurements of the levels of tyrosine hydroxylase (TH) involved in neurotransmitter synthesis in the brains of different strains of mice reveal a 50 % difference in TH activity and protein amount. However, combined neuroanatomical-enzymological analyses show that this difference is due to a 50% difference in the number of dopaminergic neurons in one strain *versus* the other. This difference is most evident in specific brain areas, namely in one area of the *substantia nigra* [28]. Similarly, *in situ* analyses of receptor distributions in different strains of mice have revealed different densities of both alpha-amino-3-hydroxy-5-methyl-4-isoxazoleproprionic acid (AMPA) and *N*-methyl-D-aspartate (NMDA) receptors in the hippocampal formation and there is no reason to doubt that similar degrees of variation occur in humans.

Many fundamental manifestations of a disease arise during early development, and hence their genesis represents an accumulation of readjustments leading to cascade effects that are manifested after birth. In autism, it is problems in embryogenesis that ultimately manifest themselves in speech abnormalities and social interactions. The initial critical lesions arise around the time of closure of the neural tube [29], giving rise to such a broad spectrum of brain damage and behavioral symptoms, that little information on the molecular origin of autism is likely to come from molecular analyses of adults. Thus while a number of autistic patients exhibit gyral anomalies, these are located in different brain regions in different individuals; and defects can not be attributed to any one brain region. Thus while "functional symptoms are similar in all subjects, the brain damage is not". In the case of Tourette's syndrome, a multitude of effects, from genetic and prenatal factors, maternal stress, perinatal factors and the development and functioning of specific circuitry, such as the cortico-striato-thalamo-cortical circuits, can all have a significant influence on the disease phenotype in later life [30, 31].

Another significant message about human diseases is that similar phenotypes can be produced by alterations in different anatomical contexts. In terms of degeneracy, there are more than 400 disorders in which hearing problems are part of the phenotype, and the same clinical end point, namely hearing impairment, can be reached from mutations in many different genes even within the same human family.

Conversely, different mutations in the same gene can give rise to different clinical phenotypes. The *RET* receptor tyrosine kinase gene, when defective, has been shown to give rise to clinically diverse conditions such as familial medullary thyroid carcinoma, multiple endocrine neoplasia 2A, MEN2B, with the additional complications of ganglioneuromas of the colon, lips, tongue, skeleton and eye, and Hirschsprung's disease, with its absence of parasympathetic innervation of the lower intestinal tract [32]. Many families often contain segregating *loci* that have large effects on the penetrance and expressivity of the phenotype, and one of the most fundamental and urgent requirements is to examine the proteomic profiles of family members with a wide variation in phenotype.

The times at which proteomic assays are conducted in the genesis and progression of a disease often determine the significance, or irrelevance, of the data. In the rat, it has been shown that the onset of noninsulin dependent diabetes mellitus is controlled by a *locus* on chromosome 1, but the progression of the hyperglycemic state is controlled by a different *locus* on chromosome 4 [33]. These different *loci*, and the networks in which they are embedded, will only be uncovered if proteomic differential display, for example, is performed at specific ages about which there is already high quality information.

Consequently, realistic analyses of proteomic networks in functional contexts mean accepting that disease phenotypes depend heavily on network variation between individuals. The challenge facing functional proteomics is to begin proteomic data capture not by avoiding this level of variation, but by harnessing it. Finally, although we have dealt exclusively with the role of genetic factors in human phenotypic perturbations, we acknowledge the importance of environmental exposures to understanding human diseases; their evaluation is simply outside the scope of this review.

8 Future experimentation

With the availability of the predicted human proteome, and with the backlog of human tissue samples that are available for analysis, how do we extract the most knowledge about human diseases from proteomic profiles and between-individual variation? Where are the experimental inroads?

The first avenue is to utilize cutting edge technologies, the second is to tailor model organism networks as closely as possible to the human situation (humanizing model organisms in terms of transgenic content), and the last is to utilize the variation between individuals to move beyond SNP analyses and to measure and interpret proteomic network fluxes. It is only by unraveling these that we can provide personalized diagnostics, and individual drug treatment for various disorders.

8.1 Technological innovations

Miniaturization and robotization, exemplified by randomly ordered fiber-optic gene arrays with molecular beacons for mutation detection [34], and massively parallel signature sequencing on microbead arrays for transcriptomics [35], are irrevocably altering the manner and speed of molecular data capture. Improved mass spectrometers

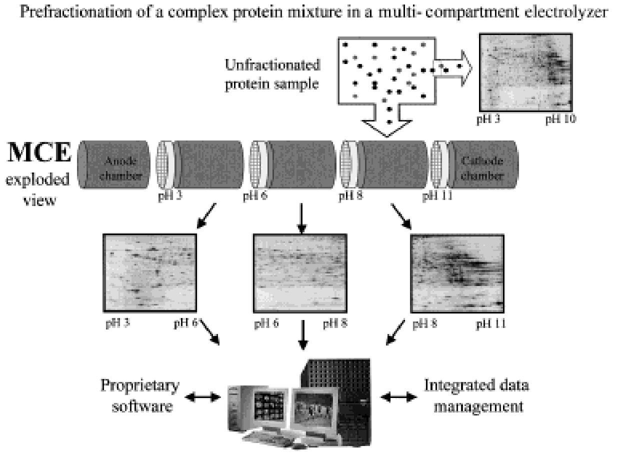
Prefractionation of a complex protein mixture in a multi- compartment electrolyzer
Unfractionated protein sample
pH 3
pH 10
MCE
exploded view
Anode chamber
Cathode chamber
pH 3
pH 6
pH 8
pH 11
pH 3
pH 6
pH 6
pH 8
pH 8
pH 11
Proprietary software
Integrated data management

Figure 1. Fractionation of a protein sample by two methodologies; conventional 2-D PAGE in a pH 3–10 range, (top right), or prefractionation in a MCE. The MCE device is shown here with only three sample chambers and two electrode chambers. The chambers are delineated by isoelectric membranes with pH values that are operator preselected to allow sample prefractionation in desired ranges; here pH 3–6, 6–8 and 8–11. The unfractionated protein sample is loaded into the desired chamber and protein separation is achieved by liquid phase IEF. During this process, proteins are trapped within particular chambers depending on their p*I* values. Following the completion of IEF, proteins from each chamber are separated by 2-D PAGE using a narrow range IPG, which is matched to the p*I* range of the fraction.

such as the MALDI QqTOF machine [36], and the potential of gel-less systems involving fourier transformed ion cyclotron resonance mass spectrometers [7], are ongoing developments whose potentials and limitations are still under evaluation. Devices such as MCE with isoelectric membranes (Fig. 1), coupled with high-throughput 2-D PAGE and MS, provide exquisite resolution in protein separation that is unmatched by other methods utilizing whole proteomes [37, 38]. Furthermore, protein-protein interaction methodologies have been scaled to a high level [39], and some can now be done on protein chips [40, 41]. In addition, combinatorial chemistry approaches [8, 9], in conjunction with systematic evolution of ligands by exponential enrichment (the SELEX technology), are producing high specificity and high affinity RNA aptamer ligands for various targets. These aptamers can be utilized for protein capture on microarrays. Finally, super sensitive biomolecular detection technologies such as MPD (Fig. 2) [42], are achieving attomole and zeptomole levels of detection, which are hundreds of times more sensitive than prior art methods, and may provide a backbone for proteomic differential display when only minute amounts of human tissue are available for analysis.

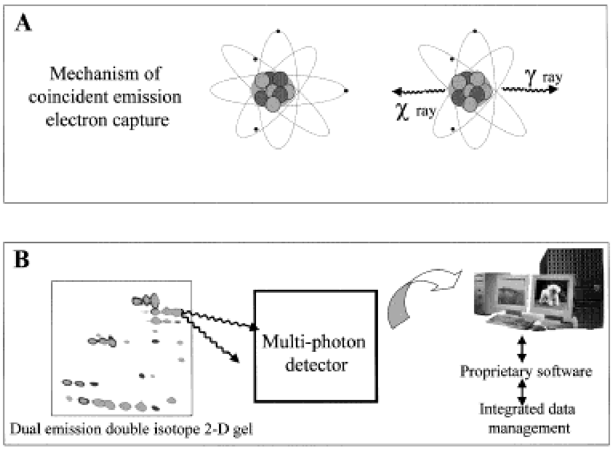
A
Mechanism of coincident emission electron capture
γ ray
χ ray
B
Multi-photon detector
Proprietary software
Integrated data management
Dual emission double isotope 2-D gel

Figure 2. A. Many isotopes which emit gamma photons have a complex radioactive decay with a near instantaneous emission of other photons. When a gamma photon is produced from the nucleus as an electron is captured, an X-ray is also produced from the disturbed electron shells. MPD uses coincident multichannel detection of the photons emitted from appropriately selected radioisotopes, together with computer controlled electronics to selectively count only those photons that are compatible with the decay signature of an operator selected radioisotope. The MPD system only reports if two photons of the expected energies are registered nearly simultaneously in opposing detectors. B. Two protein samples, control and experimental, each labeled with a different isotope, are mixed, subjected to conventional 2-D PAGE and analyzed by two "color" detection. Computer controlled electronics scan, detect and analyze the gel and integrated data management allows differential display analysis and interrogation of knowledgebases. The versatility of the proprietary detection system permits the use of over 100 different isotopes as reporter labels, providing true multicolor capability that allows simultaneous analysis of different analytes within the same gel.

While the throughput of protein chips and gel-less systems incorporating fourier transform ion cyclotron MS is increasing, these roboproteomic technologies are not yet industrialized. The surface enhanced laser desorption/ionization system (SELDI), for example, is currently restricted to the analysis of proteins that are below 30 kD in size, so a large portion of any proteomic profile remains out of reach. However, such methods are suitable for diagnostics and have provided biomarkers for ovarian and prostate cancers and for components of Alzheimer's disease [43].

The *status quo* is that robust and reliable data production for the proteomic profiles of human and mouse tissues and subcellular organelles still largely emanates from 2-

D PAGE coupled with MS [44–49]. Furthermore, it is automated multidimensional liquid chromatography (LC), electrospray ionization-mass spectroscopy (ESI-MS), and tandem mass spectroscopy (MS/MS) of proteomes that are revealing the bulk of post translational modifications, the constituents of multisubunit protein complexes, and the properties of membrane proteins, and complex protein mixtures [50–53]. Although the task of protein identification and quantification can be ramped up using isotope encoded affinity tags, there is as yet no methodology for dealing with post-translational modifications on a proteome-wide basis in a high-throughput manner.

8.2 The genesis of disease *versus* the consequences of disease

As outlined in the examples of autism and noninsulin dependent diabetes mellitus, the competitive edge will come in knowing from where to gather the data, and at what time during development. When diseases have their origins in early development, it is of limited utility to sample material such as blood or cerebrospinal fluid from adult patients if the objective is to understand the genesis of the disease. Such secondary measurements will not reflect the original causes but will instead report upon the consequences of the properties of networks that have been perturbed in embryogenesis, and reached some *quasi* equilibrium in adults.

The first requirement therefore is to produce exhaustive protein- and peptide-based 3-D *in situ* reconstructions pivotal to allowing visualization of proteomic fluxes during human organogenesis [54]. These will detail the cellular, extracellular and anatomical locations of the components of normal individual proteomic profiles. The second requirement is to distinguish between changes in proteomic profiles that remain confined to the normal anatomical real estate in which a protein is expressed *versus* gain of function and misexpression in a diseased state, where novel expression may occur in different cell types or different extracellular locations and interfere with existing networks. Without this basic knowledge, we shall continue to be working in a data mist.

8.3 Variation between individuals

We still have little idea of the global level of variation in proteomic-profiles between a sample of normal individuals, and this deficit has serious ramifications for the amelioration of most human conditions *via* pharmaceutical intervention. Humans differ greatly in their abilities to metabolize different drugs, with codeine being just one very pragmatic example. The consequences of this lack of a between-individual knowledgebase are enormous in terms of pain relief. Without such crucial data, it is not possible to assign biological significance to much of the variation currently emanating from transcriptomic or proteomic differential display analyses.

Proteomic profiles will always need to derive from individual organisms. As is clear from every case that has been examined (colon cancers, nevoid basal cell carcinomas, prostate cancers, familial medullary thyroid carcinomas, multiple and endocrine neoplasia, hearing loss, chronic inflammatory skin diseases, myotonic dystrophy, and all neuropsychiatric disorders *etc*), members of the same family carrying the same mutation can exhibit various forms of a disease, or few, or no symptoms whatsoever. To understand this variability firstly requires the harvesting of some baseline data from normal individuals, and monozygotic twins are the ideal starting point.

For example, it is only recently that the relative contributions of genetic and environmental factors on brain morphology have been quantified in the aging process using volumetric magnetic resonance imaging data and monozygotic as well as dizygotic twins [55, 56]. The data reveal that there are large genetic influences on brain cerebrospinal fluid and white matter hyperintensity volumes, the latter being more prevalent in patients with cardiovascular disease risk factors and hypertension. Excessive levels of white matter hyperintensities, for example, correlate with cognitive decline, brain atrophy, Alzheimer's and impairments in cerebral metabolism. An extension of such approaches to more biochemical levels is essential, since they will provide the most crucial information on levels of variation, particularly on the proteomic network components that are most susceptible to alteration in different genetic backgrounds, and on which components are most tightly correlated at the transcriptomic and proteomic levels. At the moment there are vanishingly few data on the extent to which proteomic networks are buffered, and hence which are particularly susceptible to direct or indirect perturbations in disease. Proteomic twin studies are likely to have significant impact in generating relevant data sets, and they remain the pre-eminent platform for differential display diagnostics.

8.4 Model organisms

Since early human development is not easily accessed, one of our choices is to extract more knowledge, more efficiently, and in a much more focussed manner, from model systems. At the functional level, the three model organisms (*Mus*, *Rattus* and *Drosophila*), will of necessity make up the main triumvirate until such time as access to human embryogenesis and foetal development is permitted, or the bottleneck is circumvented using primates, pigs and dogs. It is clear that in comparison to the mouse, the clinical responses of dogs, their disease presentations and their physiology is often closer to the human disease condition than the mouse. This is certainly the case in cancers (such as osteosarcomas), where histopathological appearance and response to therapy can be more similar between dog and human, than between human and mouse. Thus although less than 400 canine genetic disorders have been described, a majority have sufficient overlap with human disorders that they provide a convenient short circuit for analysis of the human condition [57]. Finally, the pig is a production animal and it may yet provide the significant embryological component that is missing from developmental proteomics. With a tight developmental time of approximately 114 days from fertilization until birth, with a well developed and large nervous system, and with thousands of accurately staged embryos potentially available for analysis, it could provide in-depth data for neuro-proteomic projects.

8.5 Humanized *Drosophila* and post-translational modifications

The combined power of vertebrate and invertebrate systems in providing insights into some human diseases could be galvanic, particularly when they are "humanized" by the addition of human genes to their genomes. In the case of the Machado-Joseph neurodegenerative disease (spinocerebellar ataxia type 3), progress has been slow in both humans and mice in experimentally determining how to overcome the toxic effects of nuclear inclusions that lead to neuronal death. However, the rapidity of the *Drosophila* screening methodologies in assaying for suppressors of these toxic effects, and the short transgenic turn around time of adding human genes (or parts of genes, or selected peptides) to *Drosophila*, has rapidly resulted in exquisite data. Thus co-expression of human heat shock protein 70 and human spinocerebellar ataxia 3 proteins in transgenic *Drosophila* totally suppresses the neurodegenerative phenotype. In addition, humanized flies containing the human alpha synuclein protein are providing novel insights into Parkinson's disease characterized by loss of dopaminergic neurons in the *substantia nigra*. Furthermore, other genome-wide genetic screens are rapidly uncovering different protein networks that suppress polyglutamine toxicity and other debilitating human conditions.

These data types are precisely what is required, rapid information generation on how to ameliorate a human condition. The human counterparts of the *Drosophila* suppressors that are unearthed in any screen can be further examined and their role in alleviating the condition evaluated [58–61]. Genome-wide suppressor screens of this type simply can not be done in the mouse in equivalent time frames.

Thus although the neuroanatomies of the organisms can be quite diverse, basic molecular principles can still be unearthed provided the strategies to find them are carefully chosen. The key to using model organisms is to determine *a priori* which networks to invest in, and which ones to leave well alone. Thus many of the proteins involved in vascular development in vertebrates are not present in *Drosophila* (vitronectin, fibrinogen, and von Willebrand factor), and many extracellular matrix proteins, (fibronectinelastin, tenascins, thrombospondins and osteopontin) are absent as well [62]. The universe of extracellular matrix molecules may well be sufficiently different between these two evolutionary lineages that in depth pursuit of commonalities will be difficult.

The tasks of deconvoluting post-translational modifications, such as the basic principles of heparan sulphate biosynthesis, and the components of proteoglycan function in embryogenic processes will again need to be trialled in less complex systems such as *Drosophila*, as the mammalian systems are too complex and time consuming for the requisite transgenic manipulations. *Drosophila* provides the rapidity, precision, the genetic and biological knowledgebases, mosaic technologies, sufficient overlap with some human networks, and the fully sequenced genome to carefully manipulate genomic components so that proteins are strictly targeted to the required sites at the correct times [63–67]. For example, to construct an organism that has simultaneously modified levels and precisely controlled tissue specific expression patterns of the *Notch* network, which has significant overlap with the mammalian system involved in leukemia, breast cancer, stroke and dementia, minimally requires transgenic and post-translational manipulations of *Notch*, *Delta*, *Serrate, Suppressor of Hairless*, *Enhancer of split*, *Achaete-Scute* and *Fringe* in context-dependent developmental situations. This is a multigenerational genetic undertaking that would take years in the mouse, but only months in the fly. In addition, while *Drosophila* has only single *loci* for many of these functions, mammals have at least four *Notch loci*, three *Fringe loci*, and at least five *Hairy* and *Enhancer of split* counterparts. *In vivo* manipulation of the *Drosophila* system is initially far preferable to the combinatorial transgenic nightmare that would ensue from attempting to precisely manipulate the equivalent multicomponent mouse network in a specific genetic background, and then attempting to understand it at the anatomical and phenotypic levels.

8.6 Nontraditional avenues

In addition to the above, nontraditional avenues need to be instantiated. The advent of *in vivo* imaging of the rat brain using microPET, a high resolution positron emission tomography (PET) scanner with novel detector technology provides a significant bridge for interfacing human conditions with model systems [68]. While PET is in extensive use in scanning humans with brain lesions that are the result of strokes, head injuries and surgical procedures, the implementation of microPET will facilitate investigations into animal models of epilepsy, brain disorders, and the results of transgenic modifications of ligand-receptor systems in disease and therapeutic ameliorations of disease. The ability to examine the neurotherapeutic effects of drugs, such as AMPA antagonists that protect against stroke and trauma [69], could be profitably examined *in vivo* with microPET.

Two further state of the art technologies are incoming; the 3-D *in vivo* imaging of gene expression using magnetic resonance imaging of embryos that are optically opaque [70], and implanted microdevices. Although still in their infancy, these should allow the interrogation of patterns of gene expression in a noninvasive manner. Miniaturized electronic devices implanted into mice already report upon blood pressure, heart rate and other physiological parameters in a wireless mode. The increasing sophistication of such micromonitors, particularly if tailored to proteomic measurements in transgenic or drug treated rodents, will provide a new horizon in micro engineered diagnostics and therapeutics.

Ultimately, all technologies need to complement phenotypic testing, be it by classical mutagenesis, insertional inactivation, RNA interference, protein and RNA aptamer usage, misexpression analysis in embryogenesis, and the sophisticated exploitation of multigenerational transgenic analyses in different genetic backgrounds. This is the unavoidable coalface at which all automated methods face their functional tests. Their relevance will emerge either from direct testing in transgenics or drug-perturbations in whole organisms.

9 Concluding remarks

The reality of the collision between rapid automated global methods in functional proteomics and slower higher quality data gathering is put into stark focus by the examples we have presented. In going from genomes to proteomes to anatomical proteomic profiles, the rules alter irrevocably and markedly. The pragmatism is that conventional bioinformatic methods allow a finer and finer analysis of individual components, whereas what is now more urgently required is knowledge about where and when particular proteins are expressed, and whether their perturbation in a disease state is a biologically relevant finding, or a default outcome of a pathophysiological cascade that bears little relevance to the causal underpinnings of a disease. In addition, it is still unclear how to best study several expression profiles simultaneously, how to extract functionally meaningful clusters, and how to extract clinically relevant, let alone statistically relevant, information from such comparative data sets [71, 72]. Increasingly sophisticated cluster analyses, however, will be pivotal to deconvoluting current data sets. Progress in molecular medicine will accelerate when the emphasis shifts to multitissue proteomic-profiling in epigenetic contexts, in the human case this being to embryogenesis, foetal and early adult development. Provided that a significant component of the proteomic interrogation utilizes between-individual variation, there is every expectation that by initially using specifically selected model organism systems that are most relevant to human diseases, we can track down the critical components of biological networks, which when perturbed, lead to significant causal changes in phenotype. We have an interesting set of choices. We can continue to rush headlong into disease analyses using high throughput analyses at the nucleic acid and proteomic levels and generate masses of data which are overwhelming and may, or may not, be relevant to the disease under study, or we can invest in parallel, and heavily, in the most critical aspects of the genesis of diseases and their relevant cell types. Ultimately, the acid test will be what it has always been; will the integration of the information we have discussed have beneficial effects for the patient? Provided the appropriate data are utilized, the answer is an optimistic prospectus.

We thank A. Drukier, B Herbert, T. Woodage, and N. Pakker for constructive suggestions and B Herbert for illustrations.

Received October 5, 2000

10 References

[1] www.celera.com

[2] Lockhardt, D. J., Winzeler, E. A., *Nature* 2000, *405*, 827–836.

[3] Roses, A. D., *Nature* 2000, *405*, 857–865.

[4] Risch, N. J., *Nature* 2000, *405*, 847–856.

[5] Abbott, A., *Nature* 2000, *406*, 340–342.

[6] Rohlff, C., *Electrophoresis 2000, 21, 1227–1234.*

[7] Pandey, A., Mann, M., *Nature* 2000, *405*, 837–846.

[8] Gold, L., Alper, J., *Nat. Biotechnol.* 1997, *15*, 297.

[9] Rosania, G. R., Chang, Y.-T., Perez, O., Sutherlin, D., Dong, H., Lockhard, D. J., Schultz, P. G., *Nat. Biotechnol.* 2000, *18*, 304–308.

[10] Pilarsky, C. P., Hinzmann, B., Wissman, C., Kristiansen, G., Schmitt, A. O., Kaiser, S., Rosenthal, A., *Genome Sequencing and Biology, Cold Spring Harbor* 2000, 197.

[11] Masters, J.R.W., Lakhani, S. R., *Nature* 2000, *404*, 921.

[12] Bowirrat, A., Friedland, R. P., Chapman, J., Korczyn, A. D., *Neurology* 2000, *55*, 731.

[13] Knowles, R. B., Wyart, C., Buldyrev, S. V., Cruz, L., Urbanc, B., Hasselmo, M. E., Stanley, H. E., Hyman, B. T., *Proc. Natl. Acad. Sci. USA* 1999, *96*, 5274–5279.

[14] Nimchinsky, E. A., Gilissen, E., Allman, J. M., Perl, D. P., Erwin, J. M., Hof, P. R., *Proc. Natl. Acad. Sci. USA* 1999, *96*, 5268–5273.

[15] Herkenham, M., *Neuroscience* 1987, *23*, 1–38.

[16] Strong, T. V., Tagle, D. A., Valdes, J. M., Elmer, L. W., Boehm, K., Swaroop, M., Kaatz, K. W., Collins, F. S., Albin, R. L., *Nat. Genet.* 1993, *5*, 259–265.

[17] Fukamauchi, F., Mataga, N., Wang, Y.-J., Sato, S., Yoshiki, A., Kusakabe, M., *Biochem. Biophys. Res. Commun.* 1996, *221*, 151–156.

[18] Erikson, H. P., *Nat. Genet.* 1997, *17*, 5–7.

[19] Erickson, H. P., *J. Cell Biol.* 1993, *120*, 1079–1081.

[20] Brenner, S., *Curr. Biol.* 1999, *9*, R671.

[21] Miklos, G. L. G., Rubin, G. M., *Cell* 1996, *86*, 521–529.

[22] Maleszka, R., de Couet, H. G., Miklos, G. L. G., *Proc. Natl. Acad. Sci. USA* 1998, *95*, 3731–3736.

[23] Ashburner, M., Misra, S., Roote, J., Lewis, S. E., *et al. Genetics* 1999, *153*, 179–219.

[24] http://www.jax.org/tbase

[25] Colucci-Guyon, E., Portier, M-M., Dunia, I., Paulin, D., Pournin, S., Babinet, C., *Cell* 1994, *79*, 679–694.

[26] Thatcher, J. W., Shaw, J. M., Dickinson, W. J., *Proc. Natl. Acad. Sci. USA* 1998, *95*, 253–257.

[27] Miklos, G. L. G., *Daedalus* 1998, *127*, 197–216.

[28] Ross, R. A., Judd, A. B., Pickel, V. M., Joh, T. H., Reis, D. J., *Nature* 1976, *264*, 654–656.

[29] Rodier, P. M., Ingram. J. L., Tisdale, B., Nelson, S., Romano, J., *J. Comp. Neurol.* 1996, *370*, 247–261.

[30] Leckman, J. F., Peterson, B. S., Anderson, G. M., Arnsten, A. F. T., Pauls, D. L., Cohen, D. J., *J. Child Psychol. Psychiat.* 1997, *38*, 119–142.

[31] Leckman, J. F., Cohen, D. J., *Tourettes Syndrome,* John Wiley and Sons Inc, New York, 1999, pp. 1–584.

[32] Van Heyningen, V., *Nature* 1994, *367*, 319–320.

[33] Nobrega, M. A., Jacob, H. J., *Genome Sequencing and Biology, Cold Spring Harbor* 2000, 189.

[34] Steemers, F. J., Ferguson, J. A., Walt, D. R., *Nat. Biotechnol.* 2000, *18*, 91–94.

[35] Brenner, S., Johnson, M., Bridgha, J., Golda, G., *et al. Nat. Biotechnol.* 2000, *18*, 630–634.

[36] Shevchenko, A., Loboda, A., Shevchenko, A., Ens, W., Standing, K. G., *Anal. Chem.* 2000, *72*, 2132–2141.

[37] Herbert, B., Righetti, P. G., *Electrophoresis* 2001, in press.

[38] Righetti, P. G., Barzaghi, B., Fawpel, M., *Trends Biotechnol.* 1988, *6*, 121–125, and US Patent # 4971670.

[39] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., *et al. Nature* 2000, *403*, 623–627.

[40] MacBeath, G., Schreiber, S. L., *Science* 2000, *289*, 1760–1763.

[41] Irving, R. A., Hudson, P. J., *Nat. Biotechnol.* 2000, *18*, 932–933.

[42] www.biotraces.com

[43] Senior, K., *Mol. Med. Today* 1999, *5*, 326–327.

[44] Page, M. J., Amess, B., Townsend, R. R., Parekh, R., Herath, A., Brusten, L., Zvelebil, M. J., Stein, R. C., Waterfield, M. D., Davies, S. C., O'Hare, M. J., *Proc. Natl. Acad. Sci. USA* 1999, *96*,12589–12594.

[45] Langen, H., Berndt, P., Roder, D., Cairns, N., Lubec, G., Fountoulakis, M., *Electrophoresis* 1999, *20*, 907–916.

[46] Celis, J. E., Ostergaard, M., Rasmussen, H. H., Gromov, P., *et al. Electrophoresis* 1999, *20*, 300–309.

[47] Rabilloud, T., Kieffer, S., Procaccio, V., Louwagie, M., Courchesne, P. L., Patterson, S. D, Martinez, P., Garin, J., Lunardi, J., *Electrophoresis* 1998, *19*, 1006–1014.

[48] Gauss, C., Kalkum, M., Lowe, M., Lehrach, H., Klose, J., *Electrophoresis* 1999, *20*, 575–600.

[49] Williams, K. L., *Electrophoresis* 1999, *20*, 678–688.

[50] Packer, N. H., Harrison, M. J., *Electrophoresis* 1998, *19*, 1872–1882.

[51] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., Yates, J. R. III, *Nat. Biotechnol.* 1999, *17*, 676–682.

[52] Rout, M. P., Aitchison, J. D., Suprapto, A., Hjertaas, K., Zhao, Y, Chait, B. T., *J. Cell Biol.* 2000, *148*, 635–651.

[53] Oda, Y., Huang, K., Cross, F. R., Cowburn, D., Chait, B. T., *Proc. Natl. Acad. Sci. USA* 1999, *96*, 6591–6596.

[54] Streicher, J., Donat, M. A., Strauss, B., Sporle, R., Schughart, K., Muller, G. B., *Nat. Genet.* 2000, *25*, 147–152.

[55] Carmelli, D., Swan, G. E., Reed, T., Wolf, P. A., Miller, B. L., DeCarli, C., *Neurology* 1999, *52*, 1119–1124.

[56] Carmelli, D., DeCarli, C., Swan, G. E., Jack, L. M., Reed, T., Wolf, P. A., Miller, B. L., *Stroke* 1998, *29*, 1177–1181.

[57] Ostrander, E. A., Kruglyak, L., *Genome Res.* 2000, *10*, 1271–1274.

[58] Miklos, G. L. G., Maleszka, R., *Nat. Neurosci.* 2000, *3*, 424–425.

[59] Warrick, J. M., Chan, H. Y. E, Gray-Board, G. L., Chai, Y., Paulson, H. L., Bonini, M. M., *Nat. Genet.* 1999, *23*, 425–428.

[60] Fortini, M. E., Skupski, M. P., Boguski, M. S., Hariharan, I. K., *J. Cell Biol.* 2000, *150*, F23–F30.

[61] Kazemi-Esfarjani, P., Benzer, S., *Science* 2000, *287*, 1837–1840.

[62] Hynes, R. O., Zhao, Q., *J. Cell Biol.* 2000, *150*, F89-F95

[63] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., *et al. Science* 2000, *287*, 2185–2195.

[64] Rubin, G. M., Yandell, M. D., Wortman, J. R., Miklos, G. L. G., *et al. Science* 2000, *287*, 2204–2215.

[65] Hay, B. A., Maile, R., Rubin, G. M., *Proc. Natl. Acad. Sci.* 1997, *94*, 5195–5200.

[66] Perrimon, N., *Int. J. Dev. Biol.* 1998, *42*, 243–247.

[67] Rorth, P., Zabo, K., Bailey, A., Laverty, T., Rehm, J., Rubin, G. M., Weigmann, K., Milan, M., Benes, V., Ansorge, W., Cohen, S. M., *Development* 1998, *125*, 1049–1057.

[68] Kornblum, H. I., Araujo, D. M., Annala, A. J., Tatsukawa, K. J., Phelps, M. E., Cherry, S. R., *Nat. Biotechnol.* 2000, *18*, 655–660.

[69] Turski, L., Huth, A., Sheardown, M., McDonald, F., Neuhaus, R., Schneider, H. H., Dirnagl, U., Wiegand, F., Jacobsen, P., Ottow, E., *Proc. Natl. Acad. Sci. USA* 1998, *95*, 10960–10965.

[70] Louie, A. Y., Huber, M. M., Ahrens, E. T., Rothbacher, Ut., Moats, R., Jacobs, R. E., Fraser, S. E. Meade, T. J., *Nat. Biotechnol.* 2000, *18*, 321–325.

[71] Vingron, M., Hoheisel, J., *J. Mol. Med.* 1999, *77*, 3–7.

[72] Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., *et al. Nature* 2000, *406*, 536–540.