Editorial

Protein set analyses: how could this impact the clinic?

Expert Rev. Proteomics 10(4), 305–307 (2013)

Sascha Saueru

Otto Warburg Laboratory, Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany
sauer@molgen.mpg.de

> **"Protein set analysis is a promising approach to adequately grasp biological complexity and to understand and efficiently diagnose complex diseases, monitor physiological changes and successfully develop new drugs."**

Over the last century, the life sciences have generated important insights in functional processes of life, mainly due to the broad introduction of reductionism and experimental manipulation in biology and medical research during the 20th century [1]. On the basis of this extremely successful research paradigm, amongst others, scientists rationally dissected multiple molecular mechanisms of 'living systems', identified the causes of inheritable severe disorders and developed blockbuster drugs.

For example, it is nowadays a widely known textbook knowledge that some rare mutations can cause severe monogenetic diseases due to the malfunction of key proteins. A number of simple biochemical or PCR-based gene tests can be applied to diagnose monogenic diseases. Nevertheless, recent genome-wide association studies (GWAS) revealed that for many multifactorial diseases, the disease phenotypes cannot be mono-causally derived from the action of individual genes or proteins [2]. These genetic and other functional analyses not surprisingly indicated that (slowly developing) complex diseases such as atherosclerosis are rather the result of the rising deregulation of interconnected genes or the dysfunction of molecular pathways [3]. In this context, cellular clusters of chemically modifiable proteins exert many important biological functions, closely at the solubility equilibrium in the cell.

Moreover, in line with GWAS of many diseases, recent cancer genomics studies showed that instead of mutation frequency, patterns of mutations in key oncogenes or tumor suppressor genes are specific for the disease [4]. Mutation pattern information can be used to classify disease states by applying cancer mutation databases [5]. Although a large number of mutations have been identified in cancer, the important driver mutations appear to be accumulated in fewer numbers of pathways and cellular processes, which contribute to a selective cell growth advantage.

Strikingly, even genetically identical individuals can vary in their phenotypic traits due to environmental and stochastic effects. In general, complex biological phenomena depend on molecules such as proteins, just as the meaning of this text depends on the use of letters and words. But many biological or medical phenomena we are eventually interested in are primarily based on the emerging properties and functions of structured molecular networks, influenced by environmental and stochastic factors, which cannot be sufficiently described or explained by the 'parts of the sum' of mechanistic events of few dissected biomolecules [6]. This organisational principle might just be part of the often mentioned 'complexity' in biology.

Network-based organization of biomolecules leads to evolutionary beneficial redundancy, plasticity and flexibility of physiology. Key biological functions that emerge – depending on the cellular or other biological contexts [7] – on the pathway level are more robust against environmental changes and genetic alterations than

Keywords: protein set enrichment analysis • mass spectrometry • diagnostics • drug development • disease monitoring

10.1586/14789450.2013.814883

ISSN 1478-9450

functions that would be represented by only a few proteins. Single-gene or single-protein events are potentially causal for disease when the individual effect on a disease is strong and the variance is small across individuals [8]. However, this scenario is (fortunately) the exception. Most common complex diseases seem to be the result of an unbalance of the effects of environmental factors and an inherited susceptibility leading to minor variation in the expression or the activities of many interacting gene products or proteins.

> "...many biological or medical phenomena we are eventually interested in are primarily based on the emerging properties and functions of structured molecular networks, influenced by environmental and stochastic factors."

Nevertheless, it is not surprising that molecular biologists or biochemists often favor elegant, rather simple mechanisms for the understanding of complex diseases and physiology. But this approach seems to be limited to experimentally separable aspects of biology. Moreover, many researchers hope to identify few causative or at least surrogate biomarkers, which can be of additional value for diagnosis and drug monitoring. The largely successful reductionist approach will deliver further insightful detailed information in the life sciences, in particular to understand functionalities of the so far largely uncharacterized several thousands of proteins. However, complex disease processes including environmental factors or drug treatments can interfere with many proteins or molecular networks – in many different cell types or tissues. Thus, the current inefficiencies in complex disease management and pharmacological development may argue for complementary research paradigms.

Along with the availability of new powerful data gathering and data handling tools, we are now in a good position to develop efficient models to explain non-separable properties of (molecular) pathways and biological systems [9]. Furthermore, for many diseases early diagnosis would benefit from more sophisticated biomarker analyses to provide evidence for efficient preventive intervention years before the irreversible outbreak of complex diseases. Moreover, more efficient characterization of drug candidates in early preclinical and clinical phases would result in higher success rates of drug development and treatment.

In practice, pattern recognition is a powerful and robust analytical approach and has been successfully applied for many purposes. In the proteomics field, in particular MALDI mass spectrometry, detection of mass patterns or profiles of unidentified proteins derived from whole bacterial cells has become a very popular, cost-efficient approach for routine microbial diagnostics and is now widely replacing traditional biochemical tests in the clinics [10,11]. Similar mass spectrometry-based protein pattern detection approaches were also applied for disease classification by using various body fluids to enrich in standardized procedures fragments of protease-digested proteins of human patients [12,13]. However, although useful for diagnostics, these two MALDI mass spectrometry-based molecular phenotypic approaches could not reveal much biological insight.

Biomolecule set or molecular pathway analysis is based on the idea that cellular or physiological changes manifest at the level of co-regulated, interacting or co-evolving biomolecules, rather than individually [6]. As we have shown recently, analyses of comprehensive sets of identified, well-characterized proteins can efficiently decipher functional molecular networks of complex diseases such as diet-induced insulin resistance and can help to assess drug treatments *in vivo* [14]. This protein set enrichment analysis approach [14–16] adapts the concept of gene set enrichment analysis [17], a popular bioinformatics tool to determine whether an *a priori* defined set of related biomolecules indicates statistically significant, concordant differences between various biological states (more information can be found at [101]). Gene or protein sets can be defined by the researcher, for example, by consulting various publicly available pathway databases. Using gene set enrichment analysis or protein set enrichment analysis, the whole transcriptomics or proteomics dataset, including only marginally or completely unregulated mRNAs or proteins, is used for analysis.

Using nano-liquid chromatography coupled to electrospray ionization mass spectrometry we identified and quantitatively analyzed several thousands of proteins including posttranslational modifications such as phosphosites in peripheral metabolic target tissues [14]. Applying protein set enrichment and pathway analyses, we discovered striking changes in protein pathways that indicated differential regulation of cellular and tissue homeostasis during high fat diet and antidiabetic medication. For example, in the case of the diabetes drug rosiglitazone, we could early on extract dysregulated protein pathways in the heart, even after only a few weeks of treatment and long before any typical pathological cardiovascular phenotypes could be observed. Further, using protein set analyses in conjunction with widely applied physiological assays, we could exclude the known side effects of rosiglitazone for another, new class of natural antidiabetics, the amorfrutins [18].

> "Biomolecule set or molecular pathway analysis is based on the idea that cellular or physiological changes manifest at the level of co-regulated, interacting or co-evolving biomolecules, rather than individually."

In many biological systems it seems that transcriptional (or mRNA) networks in cells are at least in part affected by stochastic processes due to Brownian motion of the mRNAs, in particular, in the case of low-abundant transcripts. In contrast, proteins tend to be expressed in larger quantities, making cellular protein networks rather non-stochastic and more robust against environmental challenges. In drug-treated or non-treated insulin-resistant obese mice, we observed only a subtle variation of individual protein expression for most (rather abundant) proteins. Interestingly, we observed striking consistence of RNA and protein expression on the pathway but much less on the individual gene-protein level [14]. These results strengthened the idea that physiological outcomes arise from the context-specific interaction of various biomolecules, which can be grouped in functionally distinct gene or protein sets. Moreover, the observed stability of regulation on the level of sets of proteins makes the presented approach largely independent from potentially non-identified proteins or still missing protein information.

Protein set analysis will strongly benefit from further instrumental improvements such as detection speed and sensitivity of applied electrospray ionization mass spectrometers. For example, modern benchtop orbitrap mass spectrometers utilizing electrospray ionization can process single-shot proteome analyses in a few hours [19], making large-scale application of protein set analyses easily doable. Using isotopically labeled human reference cells or well-controlled label-free protocols [20], protein set analysis can be extended for clinical applications such as early disease detection and treatment monitoring using biopsies or other material of human patients. The focus on the sets of proteins instead of only few biomarkers may also reduce the frequently encountered problem of interpretation of inter-patient variability.

In a nutshell, protein set analysis offers great opportunities for basic research and clinical applications [14]. This functional large-scale proteomics methodology can produce highly informative data for systems-based research. Furthermore, protein set analysis can provide powerful diagnostic read-out even at an early stage and in an unbiased way to monitor the effects of drug treatment, in the animal model or in the human patient. The mass spectrometry technologies required for protein set analysis will be mature for large-scale applications in the very near future to initiate first clinical studies.

Comprehensive quantitative protein expression and modification data will be important to model disease states and treatment regimes. Protein set analysis is a promising approach to adequately grasp biological complexity and to understand and efficiently diagnose complex diseases, monitor physiological changes and successfully develops new drugs.

Financial & competing interests disclosure

Our work is supported by the German Ministry for Education and Research (BMBF, grant number 0315082), the EU [FP7/2007-2013, under grant agreement n° 262055 (ESGI)], and the Max Planck Society. I thank C Weidner, D Meierhofer and A Freiwald for fruitful discussions and teamwork. The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

1 Weinberg R. Point: hypotheses first. *Nature* 464(7289), 678 (2010).

2 Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA* 109(4), 1193–1198 (2012).

3 Feldmann R, Fischer C, Kodelja V *et al.* Genome-wide analysis of LXRalpha activation reveals new transcriptional networks in human atherosclerotic foam cells. *Nucleic Acids Res.* 41(6), 3518–3531 (2013).

4 Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science* 339(6127), 1546–1558 (2013).

5 Forbes SA, Bindal N, Bamford S *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 39(Database issue), D945–D950 (2011).

6 Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proc. Natl Acad. Sci. USA* 104(21), 8685–8690 (2007).

7 Grueneberg DA, Degot S, Pearlberg J *et al.* Kinase requirements in human cells: I. Comparing kinase requirements across various cell types. *Proc. Natl Acad. Sci. USA* 105(43), 16472–16477 (2008).

8 Weeks DE, Lathrop GM. Polygenic disease: methods for mapping complex disease traits. *Trends Genet.* 11(12), 513–519 (1995).

9 Sauer S, Lange BM, Gobom J, Nyarsik L, Seitz H, Lehrach H. Miniaturization in functional genomics and proteomics. *Nat. Rev. Genet.* 6(6), 465–476 (2005).

10 Kliem M, Sauer S. The essence on mass spectrometry based microbial diagnostics. *Curr. Opin. Microbiol.* 15(3), 397–402 (2012).

11 Sauer S, Freiwald A, Maier T *et al.* Classification and identification of bacteria by mass spectrometry and computational analysis. *PLoS ONE* 3(7), e2843 (2008).

12 Freiwald A, Mao L, Kodelja V *et al.* Differential analysis of Crohn's disease and ulcerative colitis by mass spectrometry. *Inflamm. Bowel Dis.* 17(4), 1051–1052 (2011).

13 Baumann S, Ceglarek U, Fiedler GM, Lembcke J, Leichtle A, Thiery J. Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clin. Chem.* 51(6), 973–980 (2005).

14 Meierhofer D, Weidner C, Hartmann L *et al.* Protein sets define disease states and predict *in vivo* effects of drug treatment. *Mol. Cell Proteomics* 12(7), 1965–1979 (2013).

15 Isserlin R, Merico D, Alikhani-Koupaei R, Gramolini A, Bader GD, Emili A. Pathway analysis of dilated cardiomyopathy using global proteomic profiling and enrichment maps. *Proteomics* 10(6), 1316–1327 (2010).

16 Cha S, Imielinski MB, Rejtar T *et al.* *In situ* proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology. *Mol. Cell Proteomics* 9(11), 2529–2544 (2010).

17 Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* 102(43), 15545–15550 (2005).

18 Weidner C, de Groot JC, Prasad A *et al.* Amorfrutins are potent antidiabetic dietary natural products. *Proc. Natl Acad. Sci. USA* 109(19), 7257–7262 (2012).

19 Michalski A, Damoc E, Hauschild JP *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell Proteomics* 10(9), M111.011015 (2011).

20 Mann M, Kelleher NL. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl Acad. Sci. USA* 105(47), 18132–18138 (2008).

Website

101 Gene Set Enrichment Analysis. http://www.broadinstitute.org/gsea/index.jsp