Journal of Applied Biosciences 30: 1861 - 1865
ISSN 1997–5902

Utilization of proteins and nucleic acids in the study of gene function: a comparative review

-

Review paper-

Mwololo JK[1*], Karaya HG[2], Munyua JK[3], Muturi PW[1], Munyiri SW[1],
[1]Makerere University, Faculty of Agriculture, Crop Science Department, P.O. Box 7062 Kampala, Uganda; [2]International Maize and Wheat Improvement Centre (CIMMYT), P.O. Box 1041-00621, Nairobi, Kenya; [3]University of Nairobi, P.O. Box 30197-00100 Nairobi, Kenya.

***Corresponding author email:** mwololojames@yahoo.com; Cell phone +254720576335
Original submitted in 15th March 2010. Published online at www.biosciences.elewa.org on June 9, 2010.

ABSTRACT
Proteomics is one of the fastest growing areas in areas of research, largely because the global-scale analysis of proteins is expected to yield more direct understanding of function and regulation than analysis of genes. Protein structure characterizes its function and a protein sequence that relates to a known structure forms a basis for identifying gene function. Proteins are encoded by the genome (genes), and the set of proteins encoded by the genome, including the added variation of post-translational modification, constitute the proteome. The proteins are involved in nearly all metabolic activities, hence are part of the tools that make living machines work. The proteome is neither as uniform nor as static as the genome. However challenges encountered in identifying the biochemical and cellular functions of the many gene products which are currently not yet characterized has necessitated the use of the proteome. Gel electrophoresis techniques allow the separation of cellular proteins on a polymer according to their molecular weight and isoelectric point. The development of automated methods for the annotation of predicted gene products (proteins) with functional categories is becoming increasingly important. Compared to the study of the genetic code, proteomics may allow greater understanding of the complexity of life and the process of evolution due to the large number of proteins that can be produced by an individual organism. The measurable changes in protein profiles are also being used in diagnosis of emerging diseases. A major challenge to proteomics is that proteins are dynamic and interacting molecules, and their variability can complicate detailed studies on gene function. Nevertheless, measuring the intermediate step between genes and proteins i.e. the messenger RNA (mRNA) or the transcriptome bridges the gap between the genetic code and the functional molecules that regulate cell functions. This review examines protein amenability to prediction of gene function and the potential of proteomics in biological research.

Key words: Protein, proteome, genome, annotation, transcriptome, genetic code

INTRODUCTION

A protein is a complex polymer which is made up of amino acids. Relating protein sequence to a known structure paves way for identification of gene function. The major role of protein structure is to characterize function and there are four basic types of proteins, i.e. (i) primary, which is a sequence of amino acids; (ii) secondary, which has local folding patterns (alpha helix, beta sheet); (iii) tertiary, which is a complete 3 dimensional fold and (iv) quaternary, which is the functional state of a protein. At the quaternary structure the protein can be further organised into motifs and domains. A domain is a part of protein sequence and structure that can evolve, function, and exist independently of the rest of the protein chain. Each domain forms a compact three-dimensional structure and often can be independently stable and folded, containing an individual hydrophobic core built from secondary structural units connected by loop regions, and having a specific function. The set of proteins encoded by the genome, including the added variation of post-translational modification constitute the proteome.

Proteins and DNA

Proteins are encoded by the genome (genes), and are involved in cellular metabolic activities (Pandey & Mann, 2000). Three major characteristics of the DNA molecule make it an extremely useful tool for species identification. First, DNA is an extremely stable and long-living biological molecule that can be recovered from biological material that has been under stress conditions (e.g. processed food products, coprolites, mummified plant tissues and blood stains, among others). Second, DNA is found in all biological tissues or fluids with nucleated cells (or non-nucleated cells with plastids and/or mitochondria), enabling its analysis from almost all kinds of biological substrates. In addition, DNA can provide more information than proteins due to the degeneracy of the genetic code and the presence of large non-coding stretches.

The proteome is neither as uniform nor as static as the genome. Whereas the genetic code is made of four nucleotides and the sequence of these nucleotides is identical in every cell of an organism, proteins are built from 20 different amino acids, and post-translational modification adds other chemical constituents to these molecules, including sugars, fats, phosphates, and even other proteins. In addition, proteins come in different isoforms, are churned out through metabolic and degradative pathways, are alternatively spliced and often link with one another to form complexes of multiple proteins. Moreover, the set of proteins produced by a cell varies depending on cell type, cell shape, cell function, what tissue the cell resides in, and what signals the cell receives from its environment, in addition to the developmental stage of the cell.

Use of proteome in gene function studies

In bioinformatics analyses of whole genome sequences limitations have been encountered in identifying the biochemical and cellular functions of many gene products. To date, many remain uncharacterized. Full realization of the information encoded in genome sequences requires knowledge of the three-dimensional (3-D) structures of gene products, since it is at this level that gene function is expressed. Protein 3-D structure has traditionally provided the basis for understanding functions that have already been determined biochemically and for applications in medicine and biotechnology such as protein engineering and structure-based drug design. Currently, however, with increased throughput, it also offers a route to the discovery of function for the many gene products that are yet to be characterized (Teichmann *et al*., 2001).

Two-dimensional gel electrophoresis allows the separation of cellular proteins on a polymer according to their molecular weight in one dimension and according to their isoelectric point in the second dimension and this technique allows the quantification of cellular proteins. The gene

expression monitoring and comparative studies between the gene and the product (protein) provide invaluable information about the cellular function of proteins whose function can either be known or unknown. Knowledge on the function of a protein will simplify the assigning of function to specific genes. Another set of proteomic technologies allows the identification of interactions between proteins on one hand and between proteins and DNA domains on the other. Presently, efforts in proteomic technologies are addressing the key limitations of the narrow spectrum and quantification of the protein properties that can be monitored simultaneously. The development of automated methods for the annotation of predicted gene products (proteins) with functional categories is becoming increasingly important, improving presentation of genome sequences and genome annotations to biologists in a useful way. Thus, many systems to perform protein functional annotation have been developed that employ various sources of protein information as features, including protein functional sites (Jung & Thon, 2006), sequence similarity (Martin *et al*., 2004), gene expression patterns (Pavlidis *et al*., 2002), among others. Examples of these systems are: The two-dimensional gel electrophoresis (2-DGE) or liquid chromatography followed by tandem mass spectrometry is one of the methods for profiling protein expression (Agrawal et al., 2005). The 2-DGE entails the separation of complex protein mixtures by molecular charge in the first dimension and by mass in the second dimension. However, though recent advances in 2-DGE have improved resolution and reproducibility, the technique remains difficult to automate in a high-throughput setting; Newly introduced robotic liquid chromatography systems and high-resolution analysis methods such as Fourier transform mass spectrometry, which allow the detection of several thousand proteins. Other systems include affinity chromatography, yeast two hybrid techniques, fluorescence resonance energy transfer (FRET), and Surface Plasmon Resonance (SPR) used to identify protein-protein and protein-DNA binding reactions.

Given the large number of proteins that can be produced by individual organisms, proteomics may allow greater understanding of the complexity of life and the process of evolution than the study of the genetic code alone. Use of DNA to understand such complexity is not adequate because problems of redundancy, pseudogenes, transposon elements and sequencing create inaccuracies. The exons may be separated by thousands of bases which may present a problem to software application. Moreover, genes may overlap each other and appear in different reading frames and on different strands. Proteomics doesn't only reveal information about life's complexity but it also provides insight into the vibrancy of cells and their preparedness to react to induced conditions in changing environment (Adam *et al*., 2002). Cells and tissues respond to signals and changes in their environment, and changes in the proteome must mirror those responses. For example, early changes in the health of a tissue may be detectable by changes at the proteomic level. Researchers are currently taking advantage of measurable changes in protein profiles to diagnose diseases as they emerge (Adam *et al*., 2002). Moreover, the difference in profiles is robust enough to be used as a predictive diagnostic tool of such emerging diseases.

Using proteins to investigate gene function

The proteome may not tell a cell's entire story. After all, proteins are dynamic and interacting molecules, and their state of instability can make proteomic snapshots difficult at best (Schmucker *et al*., 2000). Furthermore, there are many technical challenges in characterizing molecules that cannot be easily amplified and have several post-translational modifications. Within the proteome, the many observed layers of complexity begin with an RNA processing mechanism called alternative splicing in which a single gene can produce multiple versions of a protein (Schmucker *et al*., 2000). An example is the production of neurexins in mammals whereby three genes give rise to over

1,000 distinct proteins within the mammalian brain (Adams, 2008).

Post-translational modifications are another source of protein variation. More than 200 different types of post-translational modifications are known and it is predicted that, for each gene in eukaryotes three different modified proteins with different functions are produced (Brett *et al*. 2001). Characterizing the biochemical and cellular functions of each protein and the analysis of protein regulation and its relation to other regulatory networks also poses a challenge (Paul & Michael, 2005). More than one-third of plant genes identified by genome sequencing lack any obvious function, and our understanding of the cellular and biochemical roles of the majority of proteins is quite limited.

Transcriptome (mRNA) as an indicator of gene activity

Measuring the intermediate step between genes and proteins, i.e. transcripts of messenger RNA, bridges the gap between the genetic code and the functional molecules that run cells. In multicellular organisms, nearly every cell contains the same genome and thus the same genes. However, not every gene is transcriptionally active in every cell; hence different cells show different patterns of gene expression. These variations underlie the wide range of physical, biochemical, and developmental differences seen among various plant cells and tissues and may play a role in the difference between health and disease. Thus, by collecting and comparing transcriptomes of different types of cells or tissues, researchers can gain a deeper understanding of what constitutes a specific cell type and how changes in transcriptional activity may reflect or contribute to disease.

The proportion of transcribed sequences that are non-protein-coding appears to be greater in more complex organisms. In addition, each gene may produce more than one variant of mRNA because of alternative splicing, RNA editing, or alternative transcription initiation and termination sites. Therefore, the transcriptome captures a level of complexity that the simple genome sequence (DNA) does not (Frith, 2005). The number of transcripts can be quantified to get some idea of the amount of gene activity or expression in a cell. For example, transcript information may help to reveal what genes give stem cells their unique properties of developmental plasticity and continuous growth in culture, or which particular gene expression changes are associated with a certain disease.

Furthermore, by considering the transcriptome, it is possible to generate a comprehensive picture of what genes are active at various stages of development. Since all nucleated cells share the same genetic material, what differentiates them are the specific genes that are expressed in each cell at specific times. The genes involved in tissue-specific or developmental processes traditionally have been studied by making libraries of all expressed genes for an organ or developmental stage. Complementary DNA (cDNA) libraries give a view of actively expressed genes by capitalizing on the fact that during the transcription of mRNA in eukaryotes, a poly A tail (consisting of a long sequence of adenine nucleotides) is added. This poly A tail distinguishes mRNA from other expressed RNAs and can therefore be used as a primer site for reverse transcription. To make a library of transcribed sequences, scientists isolate all the RNA from their cells of interest and use a single-stranded primer complementary to the unique poly A tail, as well as a viral enzyme called reverse transcriptase. Since they are produced from transcribed mRNA found in the nucleus, cDNA libraries contain primarily the protein-encoding regions of the genome. Once a cDNA has been at least partially sequenced, unique polymerase chain reaction (PCR) primer pairs that identify short stretches of each cDNA can be designed.

CONCLUSION

It is clear that proteins are more amenable to prediction of gene function and therefore proteomics will increase in importance since analyses of proteins is expected to yield more

direct understanding of function and regulation of genes than would be achieved through analysis of genes themselves. However, there is need to integrate genomics, transcriptomics and proteomics to facilitate understanding of normal cell development, function and response to diseases. Africa's ability to effectively use existing and emerging technologies in molecular biology will depend largely on the level of investment in building physical, human, institutional and societal capacities. More specifically, Africa's regional innovation communities will need to specifically focus on creating and reforming existing knowledge-based institutions, especially universities, to serve as centres of diffusion of new technologies into the economy. Building such critical capacities will also entail consideration of international cooperation as well as complementary reforms in the structure and conduct of international development cooperation agencies.

REFERENCES

Adam BL, 2002. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Research 62: 3609–3614.

Adams J, 2008. The proteome: discovering the structure and function of proteins. Nature Education 1(3).

Agrawal GK, Yonekura M, Iwahashi Y, Iwahashi H, Rakwal R, 2005. System, trends and perspectives of proteomics in dicot plants Part I: technologies in proteome establishment. Journal of Chromatography 815**:** 109–123

Frith MC, 2005. Genomics: The amazing complexity of the human transcriptome. European Journal of Human Genetics 13: 894–897.

Jung J. and Thon MR, 2006. Automatic annotation of protein functional class from sparse and imbalanced data sets. *Lecture Notes in Computer Science Journal* 4316: 65–77.

Martin D, Berriman M, Barton G, 2004. A new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5(1):178-184.

Nsubuga AM, Robbins MM, Roeder AD, 2004. Factors affecting the amount of genomic DNA extracted from ape faeces and the identification of an improved sample storage method. Molecular Ecology 13: 2089-2094.

Pandey A.. and Mann M, 2000. Proteomics to study genes and genomes. Nature Volume 405 pp837

Paul B. and Michael S, 2005. Prospects and challenges in proteomics. Plant Physiology 138: 560-562.

Pavlidis P, Weston J, Cai, J, Noble WS, 2002. Learning gene functional classifications from multiple data types. Journal of Computational Biology 9(2):401–411.

Rogers NL, Cole SA, Lan HC, Crossa A, Demerath EW, 2007. New saliva DNA collection method compared to buccal cell collection techniques for epidemiological studies. Journal of Human Biology; 19: 319-326.

Shiio Y, 2002. Quantitative proteomic analysis of *Myc* oncoprotein function. EMBO Journal 21: 5088–5096.

Smith LM. and Burgoyne LA,. 2004. Collecting, archiving and processing DNA from wildlife samples using FTA(R) databasing paper. BMC Ecol 4: 4.

Teichmann SA, Murzin AG, Chothia C, 2001. Determination of protein function, evolution and interactions by structural genomics. Cuerrent Opinion in Structural Biology 11:354-363. .