ISSN:0975-9646

Accomplishments & Key Challenges in Post Genome Era of Bioinformatics: An Advancement

Archana Dash, Mamata Nayak, Tripti Swarnkar, Kaberi Das
Department of Computer Applications,I.T.E.R, Shiksha 'O' Anusandhan University
Jagamohan Nagar, Bhubaneswar-751030, INDIA

***Abstract*—** Proteomics, the study of all proteins present in a biological sample, provides many challenges in data representation, integration, and interpretation. In this article, we survey experimental and computational approaches related to proteomics, focusing on the ways in which recent biological findings complicate the mapping from genes to RNA to protein. We argue that the challenges encountered in proteomics provide a valuable lesson on the complexity of life itself, as live organisms always contradict oversimplified models of biological information flow. It is an exciting new field that combines high-throughput experimental techniques and advanced algorithms to provide a global understanding of all the proteins expressed in particular cells under particular conditions. Considered broadly, proteomics includes: techniques for identifying proteins in a sample, detecting posttranslational modifications (changes to the proteins after translation), predicting the structure and function of proteins from sequence data, and integrating information about protein sequences from different databases. For most of the last century, proteins were analyzed one at a time. Now, the availability of completely sequenced genomes (including the human genome), databases of nuclear magnetic resonance (NMR) and X-ray structures of proteins, and compilations of functional information (such as posttranslational modifications) is driving the development of computational methods that can enable direct prediction of protein structure and function in silico (proteoinformatics). In this article, we outline the biological background needed to understand the interesting issues in proteomics.

Keywords— central dogma, proteomics, codons, protein protein interactions

I.INTRODUCTION

A. *The Central Dogma of Molecular Biology*

Most chemical reactions in the cell are carried out by proteins, catalysts made out of specific sequences of simple subunits called amino acids, which can be thought of as a 20-letter "alphabet." Proteins work by folding into specific three-dimensional (3-D) structures, which are largely determined by their sequences (see the article in this issue by Ison et al., "Proteins and Their Shape Strings," for a review). The information specifying the sequences of these proteins is encoded in the genome, a very long (5.4 million for the bacterium *E. coli* and 3 billion for humans) sequence of DNA subunits called *nucleotides*, which differ chemically from amino acids (nucleotides have a four-letter alphabet compared to the 20-letter alphabet of amino acids). RNA, a molecule closely related to DNA, is also important; it can act both as an information store and, in rare but increasingly important cases, as a catalyst. Crick's "The Central Dogma of Molecular Biology" [1] describes the main flow of information in the cell. Figure 1 shows this process at a very high level. First, DNA is copied into RNA by a protein called RNA polymerase in a process called *transcription*. Then, the RNA acts as a template for protein synthesis at a large RNA/protein complex called the ribosome in a process called *translation*; such RNAs are called *messenger RNAs* (mRNAs). The ribosome uses the genetic code to read three-letter words, codons, in the RNA alphabet as single letters in the protein alphabet. Both transcription and translation rely on specific initiation and termination signals (in the DNA and RNA sequence, respectively). It was once thought that each gene encoded one enzyme [2], but modern discoveries have complicated this picture significantly

B. A More Complex Picture of the Central Dogma

Recent work in biology has revealed new mechanisms in both transcription and translation that complicate and/or violate assumptions made by the "Central Dogma." Rather than a one-to-one mapping between DNA and RNA sequences, and between RNA and protein sequences, examples of many-to many mappings at each level can be found. One important step that contributes to these processes is termed *splicing*, in which certain parts of the mRNA (called *introns*) are excised before the mRNA is translated. For example, mRNAs transcribed from two different regions of the DNA can be spliced together through a process called *trans-splicing*, resulting in a single RNA that came from two genes ([3],[4]). Similarly, a single gene can produce many mRNAs through process called *alternative splicing*, in which different pieces of the initially transcribed RNA are deleted under different circumstances. For example, three genes that are involved in the sex determination pattern of *Drosophila* are *sxl*, *tra*, and *dsx*. Each of these genes produces a pre-mRNA with two possible splicing patterns, depending on whether a fly is male or female [5].Additionally, the "Central Dogma" is violated in the cases of retroviruses and retrotransposons, which can copy their RNA, back into DNA. However, proteins cannot be copied back into RNA or DNA; the genetic code is irreversible.

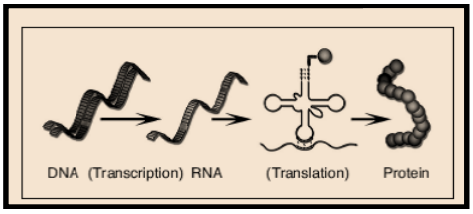

Fig. 1 Flow of information within the cell

C. Posttranslational Modifications

In addition to the complications introduced by splicing and variable transcription start and stop sites, certain amino acids in a protein can be chemically modified after translation. This process, posttranslational modification (PTM), often plays a crucial role in regulating the activity of the protein. For example, phosphorylation involves the addition or removal of phosphate groups and can activate or inactivate a protein. PTMs such as phosphorylation can also affect protein-protein interactions depending on the charge or shape of the binding surface of a protein. Another important regulatory mechanism is the binding of small molecules, producing changes in protein structure and, therefore, activity (allostery). An example of allostery is the binding of metabolic products, such as intermediates in sugar metabolism, to metabolic enzymes to modify flux through the pathway. Posttranslational modifications are also used by the cellular machinery to mark specific proteins to be broken down into their constituent amino acids. In a process called *ubiquitin- mediated degradation*, a small protein molecule called *ubiquitin* covalently attaches to a specific amino acid in a protein; when many ubiquitin molecules form a chain, the protein is targeted for degradation by a cellular structure called a *proteasome*.

D. Experimental Techniques in Proteomics

Protein structural analysis has a long history. The number of known protein sequences began to increase rapidly when Edman developed a method for sequencing proteins and peptides from the N-terminus one amino acid residue at a time, a method automated by Beckman in the 1960s [23]. Several laboratories with these automated sequencers maintained databases of the published sequences in order to identify duplicates. Dayhoff exploited these databases to provide the first analyses of evolutionary relationships among proteins [24]. Several databases once competed (reviewed in [25]), but were eventually normalized by agreement. The molecular biology revolution soon shifted the emphasis from protein sequencing to oligonucleotide sequencing, and the rate of publication of new sequences rose exponentially (since protein sequences could be inferred from the nucleotide sequences). Today, the major challenge is to combine protein and genomic databases to provide a user-friendly resource for the research community. In parallel with the development of protein sequence databases, 3-D structures of proteins were laboriously determined by X-ray crystallography. The development of the Protein Data Bank, a database of these structures [26], has led to methods for predicting structures from sequences either by homology or de novo (see article by Ison et al.). The availability of this vast array of structures (over 25,000 at the time of writing) provides a rich field for data mining as well as for theoretical and physical studies aimed at a deeper understanding of the nature of proteins. Today, proteomics encompasses several global techniques for studying large samples of proteins. Two-dimensional gel electrophoresis is a technique that separates proteins by size and pH, giving a "fingerprint" of the proteins in a sample, which can be used to identify changes in specific proteins [27]. More recently, multidimensional chromatography has been used to separate proteins, which are then analyzed by mass spectrometry (MS). Although the proteins are usually digested, the ultimate goal is to directly analyze the intact proteins by MS (referred to as *top-down proteomics*).

This will most likely require many years of methods development due to the difficulty of getting proteins into the gas phase for analysis. An emerging method referred to as *shotgun* or *bottom-up proteomics* takes advantage of the fact that peptides are much easier to analyze than whole proteins. In shotgun proteomics, all the proteins in a sample are digested into peptides, which are then separated by their chemical or physical properties. Fragmentation spectra are then collected through two rounds of mass spectrometry (tandem MS/MS), allowing identification of the corresponding peptide sequences., "Enabling Proteomics Discovery Through Visual Analysis"). However, cataloging the expressed proteins that correspond to a set of peptides remains challenging. Genome-wide, two-hybrid screens can reveal which proteins interact. Two-hybrid screens identify interactions by fusing one protein with a DNA-binding domain and a second protein with a transcription activation domain; the pairs of proteins that interact will induce expression with a reporter (a protein whose activity can be easily measured, such as the fluorescent protein GFP) [28]. A full description of a protein requires knowledge of its 3-D structure, which can assist in inferring its function. Structural proteomics is the determination of this 3-D structure. There are two basic approaches to structural proteomics. The first approach, direct determination, is conducted in the laboratory using methods such as X-ray crystallography and NMR. The major barrier to high-throughput direct determination is the preparation of suitable samples; X-ray crystallography requires crystals that diffract well, while NMR requires samples of soluble proteins smaller than 300 amino acids. Although many proteins are insoluble, small proteins are abundant in most proteomes, suggesting that NMR may have an increasing role in structural proteomics. Because many proteins are evolutionarily related, direct determination of a few structures allows others to be modeled. The second approach in structural proteomics is to use purely computational methods to predict a protein structure.

E. Key Challenges in Proteomics

Although proteomics techniques are growing increasingly powerful, many challenges still need to be overcome. The articles in this issue provide interesting approaches to address many of these challenges. It is not possible in a short review to encompass all the issues to be addressed in all areas related to proteomics. Instead, we will focus on one area (shotgun proteomics) that

demonstrates how the interplay between experimental and computational improvements can drive a field. The development of new MS instrumentation with increased sensitivity, along with better ionization methods for peptides, granted the protein chemist unprecedented analytical power for tackling complex systems. The development of algorithms that could use information from the MS data to identify the peptide sequences has also been critical. MS instruments allow information about peptide fragmentation to be collected in a high-throughput, automated fashion. Since each peptide should theoretically produce a unique spectrum of ions depending on the sequence of the peptide, the peptide fragments collected provide the input for determining the sequence. After identifying all the peptide sequences in a sample, it should be possible to identify the proteins and, consequently, the biological processes that are most active. However, several challenges remain before the dream of global sample analysis can be realized.

1) Identifying the Sequence of a Peptide from Its Fragmentation Spectrum

Because each MS/MS run produces many spectra that vary in quality, there is no robust way to directly determine a peptide from each spectrum. Software such as MASCOT [30] and SEQUEST [31] compare each measured peptide spectrum to theoretical spectra generated from a protein database, assigning scores to each match. Each spectrum often matches many peptides in the peptide database. Existing algorithms have alarmingly high false-positive rates, especially when the database contains many short sequences. Mapping peptide spectra to peptide sequences remains one of the most challenging problems in the field of proteomics.

2) Identifying the Set of Proteins from a Set of Peptides

A proteomics sample typically contains thousands of proteins, which must be identified by mapping peptides identified from mass spectra onto a protein database. Unfortunately, protein databases are often highly redundant, incomplete, and/or incorrect since the deposition of protein sequences is uncontrolled. Therefore, each identified peptide is mapped to many redundant entries, causing errors when few peptides are recovered per protein. Furthermore, the raw sequencing data can often fit more than one sequence because complete coverage of the peptide sequence may not be achieved. Statistical methods for evaluating the search results are urgently needed.

3). Resolving Isoforms of RNA and Proteins

Isoforms, produced by alternative splicing or through multigene families, often have almost identical sequences. However, different isoforms can have distinct functions in cell signaling (for instance, cells can shift the isoform patterns of metabolic proteins in response to the amount of oxygen). Since shotgun proteomics cannot find every peptide information about different isoforms (or PTMs) is often unavailable. Identifying PTMs raises similar issues. PTMs such as phosphorylation and ubiquitination modify only a few residues in a protein and therefore modify only a few of its constituent peptides. However, these peptides are usually in such low abundance that they must be enriched through chemical techniques before they can be detected. These modification processes are usually reversible, so peptides often come in both modified and unmodified forms. In “Quantitative Analysis of Proteomics Using Data Mining,” Yen et al. describe a novel method for automated quantification of protein isoforms.Such quantification can be achieved by the manual analysis of mass spectrometry signals combined with a deep knowledge of biochemistry, but this process is highly labor intensive and error prone.

4). Determining the Amount of Each Protein Expressed, and Correlating These Amounts with Other Measures of Expression

Differences in cellular activity are often caused by the differences in levels of specific proteins (or modified forms of proteins), and changes in the levels of proteins in particular signaling pathways often illuminate biological responses to specific conditions. Traditional laboratory techniques such as Western blots, which measure the binding of an antibody to one specific protein, are extremely accurate but do not scale to large numbers of proteins. Technologies such as isotope-coded affinity tagging (ICAT) [32] that can quantify proteins are immature for large-scale experiments, and their accuracy is far below that of microarrays (which measure RNA, not protein, levels).

5) Identifying the Function of a Protein from Its Sequence

Shotgun proteomics, in particular, identifies lists of thousands of proteins present in a specific sample. Yet, the functions of many of these proteins are unknown. One approach is to characterize proteins by shared motifs, which may be related to the protein’s function or regulation. In “Functional Proteomics with Biolinguisitic Methods,” Singh et al. use an *n*-gram strategy from computational linguistics. This strategy provides a functional representation of motifs in the sequences, transforming the protein sequence into a functional space where these representations may be compared using quantitative methods. An important aspect of protein functional identification is comparing the sequences as strings. In “Optimization Techniques for String Selection and Comparison Problems in Genomics,” Meneses et al. address two problems of suboptimal matching between two strings: the farthest string problem (FSP), which identifies the most distant string from a set of strings, and the related far from most string problem (FFMSP). These problems fall into a difficult category of problems known as NP-hard, which do not have polynomial-time solutions, and only approximate solutions are practical. When comparing two strings, one needs first to decide on the metric (distance) to be used to determine the difference between

the two strings. One frequently used distance is the edit distance, which is the number of editing operations (like substitutions, deletions, and insertions) needed to transform one string into the other.

6) Determining the Structure of a Protein from Its Sequence

As discussed above, finding the 3-D structure of a protein given only its sequence remains an extremely challenging problem. One approach is to use data representations that reduce the information content, making structural computations more tractable by reducing the search space; yet, it can be hard to recover molecular details from such representations. Many structure prediction methods generate thousands of possible structures for a target amino acid sequence in a reduced-content representation, but these must then be expanded and adjusted to approximate a real molecule. However, the expansion steps are computationally difficult and expensive and may produce unrealistic results. In this issue, Ison et al. describe a method to reconstruct difficult turn regions of proteins by referring to information about local conformational similarities found in all unrelated proteins. Refinement of possible structures and may also be used for comparing protein structures and searching structure databases. An initial step towards determining the 3-D structure of a protein is determining the 2-D structure, i.e., the pattern of helices and sheets.

7) Finding Protein-Protein Interactions

Studying individual proteins is only the first step. To understand the cell's function, we must understand how proteins interact with one another. In "Data Mining in Protein Interactomics," Chen et al. provide an overview of the process of collecting, analyzing, and visualizing protein protein interactions obtained from yeast two-hybrid assays. In particular, they address data representation issues for multiway interactions and interactions at different levels of abstraction as well as methods for resolving ambiguities and inaccuracies in the database. Providing visualization tools that allow the rapid manual curation of large interaction data sets is a major challenge. Also machine-learning techniques can also be applied to develop a probabilistic model, the hierarchical aspect model, for learning and predicting protein-protein interactions. The key feature of this model is using existing knowledge about proteins, such as functional classes, as latent variables of protein-protein interactions. In this model, clustering these latent variables is further performed by the other latent variable. This model enables the prediction of new protein-protein interactions.

II. CONCLUSIONS

Proteomics is an exceptionally powerful technique that allows many questions to be asked about a particular biological sample. However, several advances would allow research questions to be answered far more accurately and efficiently. For example, proteomics currently depends on many user involvements.Significant efforts in process and data modeling is needed to structure data collection and databases in a way that enables the most compelling questions to be asked. Similarly, advances in instrumentation could greatly improve the efficiency and quality of the mass spectra over today's standards. Finally, algorithms for matching spectra to sequences and for resolving the expression of different RNA and protein isoforms need to be completely overhauled to keep up with the massive amount of data. With these advances, we would be able to convert the large data sets being generated now into knowledge about biological systems and processes. Understanding even a single gene and all its products is a monumental task, especially when considering all the cells, developmental variations, disease forms, and isoforms. Global profiling is just beginning to scratch the surface. The most comprehensive study to date (ours) has sampled only an estimated 11% of the peptides in the soluble extract, which itself contains only 40%-50% of the number of total open reading frames in the cell (estimated at 12,000–15,000 based on DNA arrays). Top-down approaches that begin with whole proteins are better (but more technically demanding), and studies to date have surveyed no more than 100 proteins. Studies of protein complexes are also providing interesting but puzzling data because there is so little overlap between comparable systems. The problems in proteomics should be a cautionary lesson, indicating that computational and experimental scientists engaged in this mighty enterprise need to develop new methods of validating, processing, and interpreting data. However, the problems encountered also provide insight into the complexity of life itself.

REFERENCES

[1] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp.561–563, 1970.

[2] G. Beadle and E. Tautum, "Genetic control of biochemical reactions in Neurospora," *Proc. Nat. Acad. Sci.*, vol. 27, no. 11 pp. 499–506, 1941.

[3] D. Solnick, "Trans splicing of mRNA precursors," *Cell.*, vol. 42, no. 1, pp.157–164, 1985.

[4] S. Binder, A. Marchfelder, A. Brennicke, and B. Wissinger, "RNA editing in trans-splicing intron sequences of nad2 mRNAs in Oenothera mitochondria," *J. Biol. Chem.*, vol. 267, no. 11, pp. 7615–7623, 1992.

[5] R.N. Nagoshi, M. McKeown, K.C. Burtis, J.M. Belote, and B.S. Baker, "The control of alternative splicing at genes regulating sexual differentiation in D. melanogaster," *Cell.*, vol. 53, no. 2, pp. 229–236, 1988.

[6] D. Schmucker, J.C. Clemens, H. Shu, C.A. Worby, J. Xiao, M. Muda, J.E. Dixon, and S.L. Zipursky, "Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity," *Cell.*, vol. 101, no. 6, pp. 671–684, 2000.

[7] X. Qu, Y. Qi, and B. Qi, "Generation of multiple mRNA transcripts from the novel human apoptosis-inducing gene hap by alternative polyadenylation utilization and the translational activation function of 3_ untranslated region," *Arch. Biochem. Biophys.*, vol. 400, no. 2, pp. 233–233, 2002.

[8] C. Touriol, S. Bornes, S. Bonnal, S. Audigier, H. Prats, A.C. Prats, and S. Vagner, "Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons," *Biol. Cell.*, vol. 95, no. 3–4, pp. 169–178, 2003.

[9] G.E. Tennyson, C.A. Sabatos, K. Higuchi, N. Meglin, and H.B. Brewer, Jr., "Expression of apolipoprotein B mRNAs encoding higher—and lower-molecular weight isoproteins in rat liver and intestine," *Proc. Nat. Acad. Sci. USA*, vol. 86, no. 2, pp. 500–504, 1989.

[10] L. Hong and R.B. Hallick, "Gene structure and expression of a novel Euglena gracilis chloroplast operon encoding cytochrome b6 and the beta and epsilon subunits of the H(+)-ATP synthase complex," *Curr. Genet.*, vol. 25, no. 3, pp. 270–281, 1994.

[11] S. Karlin, C. Chen, A.J. Gentles, and M. Cleary, "Associations between human disease genes and overlapping gene groups and multiple amino acid runs,"*Proc. Nat. Acad. Sci. USA*, vol. 99, no. 26, pp. 17008–17013, 2002.

[12] R. Dorn, G. Reuter, and A. Loewendorf, "Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila," *Proc. Nat. Acad.Sci. USA*, vol. 98, no. 17, pp. 9724–9729, 2001.

[13] C.J. Saris, J. Domen, and A. Berns, "The pim-1 oncogene encodes two related protein-serine/threonine kinases by alternative initiation at AUG and CUG," *Embo.J.*, vol. 10, no. 3, pp. 655–564, 1991.

[14] G. Akiri, D. Nahari, Y. Finkelstein, S.Y. Le, O. Elroy-Stein, and B.Z. Levi, "Regulation of vascular endothelial growth factor (VEGF) expression is mediated by internal initiation of translation and alternative initiation of transcription," *Oncogene*, vol. 17, no. 2, pp. 227–236, 1998.

[15] Y. Hu, Z. Zhou, C. Xu, Q. Shang, Y.D. Zhang, and Y.L. Zhang, "Androgen down-regulated and region-specific expression of germ cell nuclear factor in mouse epididymis," *Endocrinol.*, vol. 144, no. 4, pp. 1612–1619, 2003.

[16] T. Jacks, M.D. Power, F.R. Masiarz, P.A. Luciw, P.J. Barr, and H.E. Varmus, "Characterization of ribosomal frameshifting in HIV-1 gag-pol expression," *Nature*, vol. 331, no. 6153, pp. 280–287, 1988.

[17] R. Seger, N.G. Ahn, T.G. Boulton, G.D. Yancopoulos, N. Panayotatos, E. Radziejewska, L. Ericsson, R.L. Bratlien, M.H. Cobb, and E.G. Krebs,

[18] J. Ping, J.F. Schildbach, S.Y. Shaw, T. Quertermous, J. Novotny,R. Bruccoleri, and M.N. Margolies, "Effect of heavy chain signal peptide mutations and NH2-terminal chain length on binding of anti-digoxin antibodies," *J. Biol. Chem.*, vol. 268, no. 31, pp. 23000–23007, 1993.

[19] X. Duan, F.S. Gimble, and F.A. Quiocho, "Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity," *Cell.*, vol. 89, no. 4, pp. 555–564, 1997.

[20] S. Lindquist, "Mad cows meet psi-chotic yeast: The expansion of the prion hypothesis," *Cell.*, vol. 89, no. 4, pp. 495–498, 1997.

[21] D.M. Carrington, A. Auffret, and D.E. Hanke, "Polypeptide ligation occurs during post-translational modification of concanavalin A," *Nature*, vol. 313, no. 5997, pp. 64–67, 1985.

[22] E. Latres, D.S. Chiaur, and M. Pagano, "The human F box protein beta-Trcp associates with the Cul1/Skp1 complex and regulates the stability of beta-catenin," *Oncogene*, vol. 18, no. 4, pp. 849–854, 1999.

[23] C.I. Branden, "Protein chemistry. Founding fathers and families," *Nature*, vol. 346, no. 6285, pp. 607–608, 1990.

[24] L.T. Hunt and M.O. Dayhoff, "Table of abnormal human globins," *Ann. N. Y. Acad. Sci.*, vol. 241, pp. 722–735, 1974.

[25] B.C. Orcutt, D.G. George, and M.O. Dayhoff, "Protein and nucleic acid sequence database systems," *Annu. Rev. Biophys. Bioeng.*, vol. 12, pp. 419–441, 1983.

[26] F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.F. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank. A computer-based archival file for macromolecular structures," *Eur. J. Biochem.*, vol. 80, no. 3, pp. 319–324, 1977.

[27] J.I. Garrels, "Quantitative two-dimensional gel electrophoresis of proteins," *Methods Enzymol*, vol. 100, pp. 411–423, 1983.

[28] S. Fields and O. Song, "A novel genetic system to detect protein-protein interactions," *Nature*, vol. 340, no. 6230, pp. 245–246, 1989.

[29] T.I. Zarembinski, L.W. Hung, H.J. Mueller-Dieckmann, K.K. Kim, H. Yokota, R. Kim, and S.H. Kim, "Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 26, pp. 15189–15193, 2004

[30] D.N. Perkins, D.J.C. Pappin, D.M. Creagy, and J.S. Cottrell, "Probabilitybased protein identification by sesarching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.

[31] J.K. Eng, A.L. McCormack, and J.R. Yates III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Am. Soc. Mass Spectrometry*, vol. 5, no. 11, pp. 976–989, 1994.

[32] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.