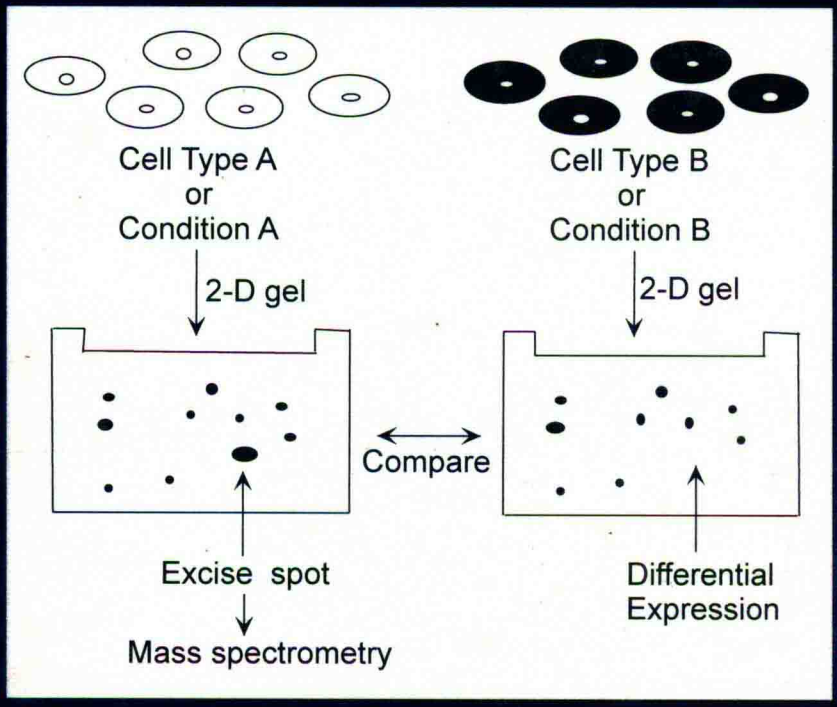
PROTEOMICS

by
Timothy Palzkill



Kluwer Academic Publishers

PROTEOMICS

by

Timothy Palzkill
Baylor College of Medicine

KLUWER ACADEMIC PUBLISHERS
Boston / Dordrecht / London

Distributors for North, Central and South America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA
Telephone (781) 871-6600
Fax (781) 681-9045
E-Mail <kluwer@wkap.com>

Distributors for all other countries:
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS
Telephone 31 78 6392 392
Fax 31 78 6546 474
E-Mail <services@wkap.nl>


Electronic Services <http://www.wkap.nl>

Library of Congress Cataloging-in-Publication Data

Palzkill, Timothy, 1961-
Proteomics/Timothy Palzkill.
p. cm.
Includes bibliographical references and index.
ISBN 0-7923-7565-3 (alk. paper)
1. Proteins—Analysis.2. Proteins—Structure.3. Protein binding.4. Gene expression.5. DNA microarrays. 6. Post-translational modification. I. Title.

QP551 .P295 2001
572'.6—dc21

2001050242

Printed on acid-free paper.
Printed in the United States of America

PROTEOMICS

PREFACE

Genome sequencing projects have produced an incredible expansion in our knowledge of the DNA sequences of a wide variety of organisms. Thus, the complete set of genes is known for many organisms. However, the function of many of the newly identified genes is not known. Nor is it known how the gene products interact to create a living organism. The field of functional genomics is an attempt to develop large-scale experimental approaches to address these questions. The approach most often associated with functional genomics is microarray hybridization. These experiments provide an assessment of RNA expression levels for all genes simultaneously. Microarrays have become an extremely popular method for understanding transcriptional regulation on a genome-wide scale. However, there is also a great deal to learn at the post-transcriptional level. Large-scale studies of post-transcriptional events are the subject of proteomics.

The name proteomics is traditionally associated with the display of large sets of proteins from a given organism or cell line on two-dimensional (2D) polyacrylamide gels. The ability to associate a spot on a 2D-gel with a known protein is used to create databases of proteins that are expressed in an organism or cell line under defined experimental conditions. This approach is complementary to the generation of databases of mRNA expression levels by microarray hybridization. The combination of technologies permits an assessment of post-transcriptional regulation and post-translation modifications. However, the field of proteomics is rapidly expanding with additional experimental approaches and this book is intended to reflect that expansion. A broader definition of proteomics is used that includes systematic experimental and computational attempts at determining protein-protein interaction maps for an entire organism.

Proteomics is an interdisciplinary science that includes biology, bioinformatics, and protein chemistry. The purpose of this book is to provide the reader with an overview of the types of questions being addressed in proteomics studies and the technologies used to address those questions. The first chapter is a concise outline of the field as it presently stands. The second chapter provides an overview of the use of 2D-gel electrophoresis and mass spectrometry to identify proteins, as well as post-translational

modifications of proteins, on a genome-wide scale. The chapter also includes an assessment of the limitations of this approach and outlines new developments in mass spectrometry that will advance future research. Chapter three describes the use of mass spectrometry to characterize the changes in protein expression profiles in different cell types or in the same cell type under different experimental conditions. The fourth chapter outlines high-throughput recombinant DNA cloning methods used to systematically clone all of the open reading frames of an organism into plasmid vectors for large-scale protein expression and functional studies such as protein-protein interactions with the two-hybrid system.

An important and growing aspect of proteomics is the attempt to generate protein-protein interaction maps for an entire genome. This information is crucial to an understanding of how genes work in concert to generate a working cell. This information, in conjunction with knowledge of transcriptional regulation obtained from microarray experiments, will provide insights into gene function. Chapter five details the experimental approaches used to generate protein-protein interaction maps including the yeast two-hybrid system, mass spectrometry and phage display. Chapter six is a summary of several computational approaches to identify protein interaction networks. Chapter seven describes attempts to create protein microarrays analogous to the DNA chips used to study RNA levels. Protein arrays hold the promise of fast, sensitive protein-protein and protein-ligand interaction mapping on a genome-wide scale. In addition, protein arrays will greatly facilitate drug discovery by allowing the rapid determination of protein targets for a prospective drug. Finally, this chapter covers efforts at determining the function of genes by the activity of the protein products. This involves the large scale cloning, expression and purification of all of the proteins of an organism. This approach has been termed biochemical genomics. Finally, chapter eight describes current limitations and possible future directions for proteomics research.

It is hoped that this book will provide the basis for understanding the field of proteomics. It is not intended to cover every aspect of the field in encyclopedic style but rather to serve as a starting point for more advanced study. Because proteomics is a young and rapidly evolving field, the best approach is to gain a general understanding of the questions and technologies involved and then pursue to the primary literature for detailed information on the latest developments.

Timothy Palzkill

CONTENTS

Chapter 1
INTRODUCTION

Whole genome sequence information is now available for many organisms. Sequence analysis of this information reveals many novel genes for which no function can be assigned. Even for well-studied model systems such as *Escherichia coli* and *Saccharomyces cerevisiae*, the specific function of approximately half of the genes is unknown. The challenge of understanding the function of each gene in the genome has led to the development of large-scale, high-throughput experimental techniques that are collectively referred to as functional genomics. These studies include systematic disruption of predicted genes, mRNA expression profiling based on microarray or DNA chip technologies, protein expression profiling using two-dimensional electrophoresis and mass spectrometry, and large scale mapping of protein-protein interactions.

Proteomics is a branch of functional genomics that has arisen in response to the inevitable question posed by the genome sequencing projects, i.e., what are the functions of all the proteins? Proteomics can be defined as the large-scale study of protein properties such as expression levels, post-translational modifications and interactions with other molecules to obtain a global view of cellular processes at the protein level. Because the tools for high-throughput DNA and RNA analysis are not available for protein analysis, the emphasis of functional genomics has been on the mRNA message. However, it is the product of the mRNA, i.e., the protein, which actually carries out the majority of the reactions of the cell. In addition, there is no *a priori* reason to expect that there will be a strict linear relationship between mRNA levels and the protein complement or 'proteome' of a cell. Proteomics is therefore a complementary approach to genomics and mRNA expression mapping using microarrays. Finally, most drug targets are proteins; therefore, methods to efficiently analyze the protein complement of cells should contribute directly to drug development.

The activity most often associated with proteomics is fractionating and visualizing large numbers of proteins from cells on two-dimensional

(2D) polyacrylamide gels. These types of experiments have been performed for more than twenty years to build databases of proteins expressed from certain cell or tissue types (Anderson and Anderson, 1996; O'Farrell, 1975). Although this remains an important component of proteomics research, the field has expanded due to the development of additional technologies. Proteomics can be broadly divided into two areas of research: (i) protein expression mapping, and (ii) protein interaction mapping.

Protein expression mapping involves the quantitative study of global changes in protein expression in cells, tissues or body fluids using 2D gel electrophoresis coupled with mass spectrometry. The identity of proteins within spots on 2D gels can be rapidly determined by in-gel proteolysis and peptide mass fingerprinting using mass spectrometry. In addition, recent developments in tandem mass spectrometry using nano-electrospray methods and enabled partial sequence information to be rapidly generated from spots on 2D gels. Thus, it is possible to generate databases of protein expression profiles for various cells and tissues (Rasmussen et al., 1996). In addition, rapid progress has been made in the identification of post-translational modifications of proteins (Oda et al., 2001; Zhou et al., 2001). This information is also being incorporated into protein expression profile databases. The aim of protein expression mapping is to compare the spectrum of proteins expressed in cells or tissues under different environmental conditions or from different disease states. Furthermore, an understanding of post-translational modifications of expressed proteins under different conditions or disease-states is sought. For clinical applications, the objective of protein expression mapping is to identify proteins that are up- or downregulated or modified in a disease-specific manner to use as diagnostic reagents or possible therapeutic targets. For basic research, the goal is to understand how the regulation of protein levels or modifications contributes to the execution and coordination of cellular processes.

Protein-protein interaction mapping involves determining, on a proteome-wide scale, the interaction partners for each of the encoded proteins of a cell or organism. The majority of the proteins within a cell are thought to work in concert with other proteins via direct physical interactions to carry out cellular processes. A great deal can be inferred about the function of a protein through knowledge of its interaction partners. For example, if a protein of unknown function is found to interact with a set of proteins known to be involved in a certain cellular process, the unknown protein can be inferred to contribute to the same process. Therefore, creation of a protein-protein interaction map of the cell would be of great value for

understanding the biology of that cell.

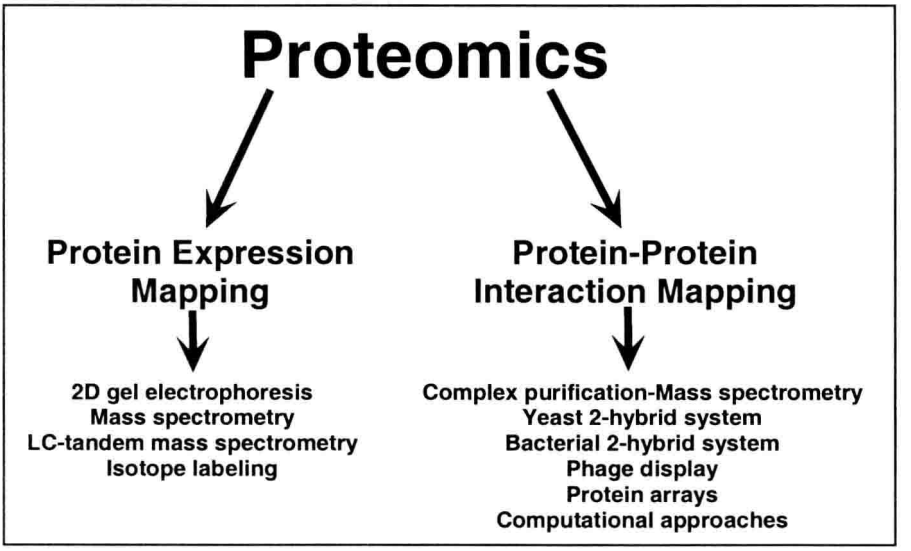


Figure 1.1. Outline of proteomics. Proteomics can be divided into two areas, protein expression mapping and protein-protein interaction mapping. The various experimental approaches used in these areas are listed. LC, liquid chromatography.

A number of technologies are available to study protein-protein interactions. For instance, the yeast two-hybrid system is an *in vivo* method that has been widely used to perform large-scale protein-protein interaction studies (Ito et al., 2001; Uetz et al., 2000). Another widely used approach has been to purify protein complexes from cells and to determine the protein components of the complex by mass spectrometry (Link et al., 1999; Rout et al., 2000). In contrast to the wide application of microarrays to study hybridization of nucleic acids, protein arrays have proved difficult to develop due the fact that proteins possess a precise (and delicate) three-dimensional structure that must be maintained on the surface of the chip. Nevertheless, there have been several reports of large-scale protein arrays and this area is under rapid development. Finally, computational approaches have been developed to predict functional interactions between proteins based on genome sequence data (Eisenberg et al., 2000). These approaches have the advantage that they can be rapidly employed to generate interaction maps for a number of organisms. The result of a computational prediction has been successfully used to guide experimental protein-protein interaction mapping for the rapid generation of a genome-wide interaction map (Newman, 2000).

Chapter 2

PROTEIN IDENTIFICATION AND ANALYSIS

One focus of proteomics is to determine the complement of proteins that are expressed in a cell and how this complement changes under different conditions. As such, the ability to accurately identify proteins on a large scale is critical to proteomics studies. Methods for protein characterization, especially mass spectrometry technologies, have greatly improved in accuracy and throughput in recent years. These new technologies have enabled the identification of hundreds to thousands of proteins from organisms and have made the characterization of entire proteomes a realistic goal.

2.1 Protein preparation and separation

Two-dimensional gel electrophoresis

Prokaryotic cells express hundreds to thousands of proteins while higher eukaryotes express thousands to tens of thousands of proteins at any given time. If these proteins are to be individually identified and characterized, they must be efficiently fractionated. One-dimensional sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) has typically been use to study protein mixtures of $\leq$100 proteins. One-dimensional electrophoresis is useful because nearly all proteins are soluble in SDS, molecules ranging from approximately 10,000 to 300,000 molecular weight can be resolved, and extremely basic or acidic proteins can be visualized. The major disadvantage to one-dimensional gels is that they are not suitable for complex mixtures such as proteins from whole cell lysates.

Two-dimensional separation (2D) involves first separating proteins based on their isoelectric point (pI) using isoelectric focusing (IEF). The isoelectric point is the pH at which there is no net electric charge on a protein. IEF is an electrophoretic technique whereby proteins are separated in a pH gradient. An electric field is applied to the gradient and proteins migrate to the position in the pH gradient equivalent to the pI (Fig. 2.1).

Because the pI of a protein is based on its amino acid sequence, this technique has good resolving power. The resolution can be adjusted further by changing the range of the pH gradient. The use of immobilized pH gradient (IPG) strips has enabled reproducible micropreparative fractionation of protein samples, which is not consistently possible when ampholytes are used in the first dimension (Gorg et al., 2000).

The second step in 2D electrophoresis is to separate proteins based on molecular weight using SDS-PAGE. Individual proteins are then visualized by Coomassie or silver staining techniques or by autoradiography. Because 2D gel electrophoresis separate proteins based on independent physical characteristics, it is a powerful means to resolve complex mixtures proteins (Fig. 2.1). Modern large-gel formats are reproducible and are the most common method for protein separation in proteomic studies.

Limitations of two-dimensional gel electrophoresis

Despite their excellent resolving power, 2D gels are limited in several respects. One problem is the sensitivity and reproducibility of detection. Proteins are expressed in cells over a wide dynamic range of concentrations but the detection range of the Coomassie or silver staining methods is limited. For example, it has been shown by silver staining of 2D gels of protein lysates from *Saccharomyces cerevisiae* that only abundant proteins are identified (Gygi et al., 2000). Even with high sample loads and extended electrophoretic separation, medium to low abundance yeast proteins are not identified. Because these proteins are encoded by approximately 50% of the yeast genes, this observation suggests 2D gel analysis as a technique for proteome characterization is inadequate (Gygi et al., 2000). These results illustrate that, despite the wide use of silver staining, the method has several drawbacks including (i) poor reproducibility, (ii) limited dynamic range, and importantly, (iii) the fact that certain proteins stain poorly or not at all.

The development of fluorescent dyes to visualize proteins from 2D gels may increase the sensitivity and reproducibility of the technique (Steinberg et al., 1996). Staining with dyes such as SYPRO Orange or SYPRO Red is noncovalent and can be performed in a simple one-step procedure after the electrophoretic steps. These dyes bind to the SDS moiety surrounding proteins and therefore show little protein-specific variability (Gorg et al., 2000; Steinberg et al., 1996). Consequently, the dyes provide more uniform staining of proteins and thus reduce protein specific detection artifacts. The detection limit of these fluorescent dyes is in the range of 1-2 nanograms of protein per spot, which is slightly less sensitive than silver staining but, in contrast to silver staining, the linear range of fluorescent staining is over three orders of magnitude. Therefore, staining with

fluorescent dyes holds promise for quantitating the amount of protein in a spot from cells grown under different conditions. Staining proteins with fluorescent dyes, however, does not completely solve the detection problems of 2D gels in that highly expressed proteins can frequently obscure the visualization of proteins expressed at low levels (Gygi et al., 2000).

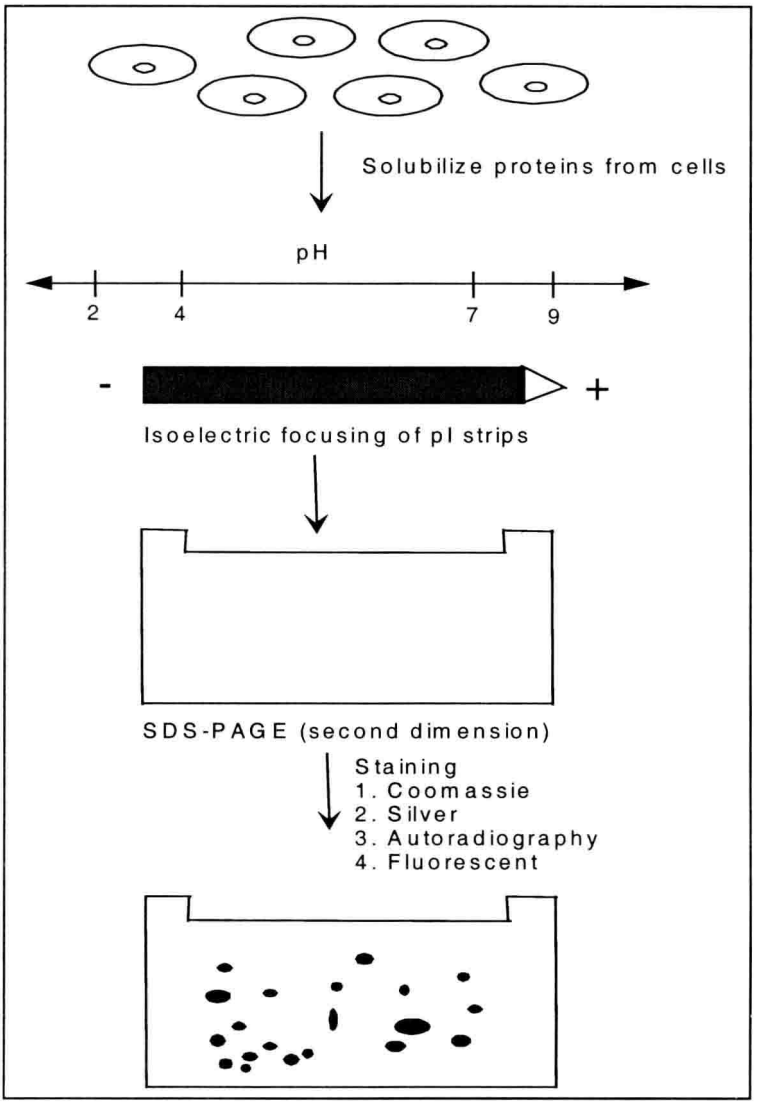


Figure 2.1. Schematic illustration of two-dimensional gel electrophoresis. Proteins are extracted from the organism of interest and solubilized. The first dimension separates proteins based on isoelectric point. The pI strip is reduced and alkylated and applied to an SDS-PAGE gel for separation by molecular weight. Proteins can be visualized using a number of staining techniques.

Another limitation of 2D gels is that membrane proteins are underrepresented. Because membrane proteins account for approximately 30% of total proteins (Wallin and Von Heijne, 1998), this is a serious problem for characterization of the proteome. The relative lack of membrane proteins resolvable on 2D gels can be attributed to three main factors: (i) they are not abundant, and therefore are difficult to detect by standard staining techniques, (ii) they often possess alkaline pI values, which make them difficult to resolve on the pH gradients most often used for isolelectric focusing, and (iii) the most important reason for under representation may be that membrane proteins are poorly soluble in the aqueous media used for isoelectric focusing (Santoni et al., 2000). Membrane proteins are designed to be soluble in lipid bilayers and are therefore difficult to solubilize in water-based solutions.

If membrane proteins are to be accurately represented, solutions to the three problems listed above are necessary. New staining techniques such as the fluorescent dyes, as well as methods that allow the loading of milligram quantities of protein onto 2D gels address the problem of abundance (Rabilloud et al., 1994). In addition, proteins with alkaline pI values can be more efficiently separated with new, wide pH range gradients (Gorg et al., 2000). In contrast, problems related to hydrophobicity of membrane proteins are more difficult to solve and progress in this area has been slow. The development of new organic solvents and detergents for the solubilization of membrane proteins is needed.

Another difficulty with 2D gel separations is posttranslational and proteolytic modifications. Although the identification of posttranslational modifications is an important aspect of detailed characterization of the proteome, it can create problems for protein identification. For example, proteolytic degradation of the sample can result in the same protein appearing at several locations on a gel. If this is an abundant protein, it can further obscure low abundance proteins. In addition, the biological significance of proteolytic digestion of a protein is difficult to assess. Posttranslational modifications such as phosphorylation or glycosylation can also place a protein at multiple positions on a gel. However, as described below, rapid progress has been made in the identification of such modifications.

Reducing complexity: Protein fractionation prior to electrophoresis

Because of the difficulties in abundance and compatibility described above, fractionation steps are often performed on protein mixtures prior to 2D gel separation to reduce the complexity of the mixtures. Prefractionation of proteins can be achieved by (i) isolation of cell compartments such as the plasma membrane or organelles such as mitochondria or nuclei, (ii) by

sequential extraction procedures with alternative solubilization capacities such as aqueous buffers versus detergents, or (iii) by fractionation methods such as free flow electrophoresis or chromatography.

The isolation of cell compartments or organelles not only provides a less complex protein mixture for 2D separations, but it also is a means to determine the cellular localization of proteins. Knowledge of cellular location can be an important clue as to the function of a protein. Numerous protocols are available for fractionation of cellular components and the details are dependent on the organism under study. Identification of a protein in a particular compartment can be confirmed by fusing the protein to a readily visible tag such as the green fluorescent protein and determining its cellular location by fluorescence microscopy and imaging (Pandey and Mann, 2000). For example, in a recent study, nuclei from mouse liver cells were isolated and specific nuclear structures named interchromatin granule clusters (IGCs) were purified (Mintz et al., 1999). The purified IGCs were then analyzed by 2D gel electrophoresis and peptide sequencing and mass spectrometry were used to identify the proteins in this structure. Seventeen proteins of unknown function were found among the IGC proteins. The tagging of several proteins with yellow fluorescent protein allowed localization of proteins to the nucleus. Similarly, the proteins of the chloroplast of the garden pea have been catalogued using a related experimental approach (Peltier et al., 2000). By using a combination of methods including chloroplast purification, solubilization of proteins, two-dimensional gel electrophoresis and mass spectrometry, a total of 200 chloroplast proteins were identified in the luminal space and periphery of the chloroplast thylakoid membrane (Peltier et al., 2000). The use of this approach to study the proteomes of organelles and other cellular structures is likely to increase in popularity.

The sequential extraction of protein samples with buffers of increasing solubilizing capacity is another means of fractionating samples. This could involve, for example, an initial extraction with an aqueous buffer followed by an extraction with an organic solvent such as methanol followed by a final extraction with a detergent (Gorg et al., 2000). Such an approach may be useful to fractionate soluble proteins from peripheral membrane proteins and peripheral membrane proteins, in turn, from integral membrane proteins (Santoni et al., 2000). Because membrane proteins and peripheral membrane proteins are poorly soluble in aqueous buffers and may only be partially soluble in organic solvents and detergents, it is important to reduce the complexity of the protein lysate to enrich and concentrate these proteins for subsequent analysis.

A number of affinity-based or chromatography methods have been used to prefractionate protein samples for 2D electrophoresis. For example, proteins of low abundance can be enriched from crude lysates by affinity-

based protein purification strategies, such as the use of an antibody specific to the protein(s) of interest. Another antibody-based approach involves immunoprecipitation of phosphorylated proteins using an anti-phosphotyrosine or anti-phosphoserine antibody (Pandey and Mann, 2000). Alternatively, or in addition to immunoprecipitation, phosphorylated proteins can be affinity purified by immobilized metal affinity chromatography (Ahn and Resing, 2001). After affinity purification, the phosphoproteins can be separated by one or two-dimensional electrophoresis and identified by mass spectrometry as described below. This is an efficient means of identifying post-translational modifications. Other chromatography strategies are less specific, such as the fractionation of proteins by charge or hydrophobicity. The objective in these cases is not to identify highly charged or hydrophobic proteins per se, but rather to provide a reproducible means of reducing the complexity of a whole cell lysate and at the same time concentrating the proteins that are fractionated. By generating a number of 2D protein gel images following a series of general, reproducible chromatographic separations, it may be possible to visualize a large fraction of the proteome of an organism of interest (Fig. 2.2).

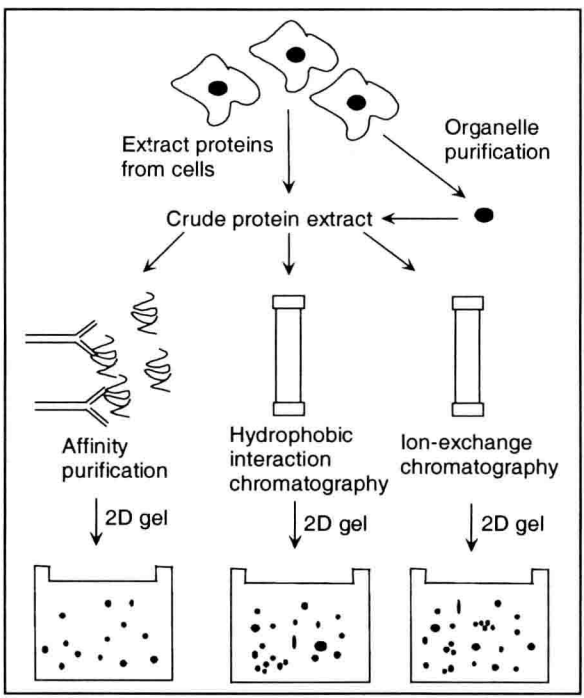


Figure 2.2. Fractionation of protein extracts before 2D gel electrophoresis. Crude lysates can be fractionated by affinity purification or by a number of chromatographic techniques. In addition, organelles or other cellular structures can be purified and lysates from these organelles can be fractionated or separated directly on 2D gels. By repeating this procedure using a number of conditions it may be possible to visualize a large fraction of a cell's proteome.