HUMAN CELLULAR PROTEIN PATTERNS AND THEIR LINK TO GENOME DNA MAPPING AND SEQUENCING DATA: TOWARDS AN INTEGRATED APPROACH TO THE STUDY OF GENE EXPRESSION

Julio E. Celis,* Hanne H. Rasmussen,* Henrik Leffers,* Peder Madsen,* Bent Honoré,* Kurt Dejgaard,* Paul Gromov,* Eydfinnur Olsen,* Hans J. Hoffmann,* Morten Nielsen,* Borbala Gesser,* Magda Puype,† Josef Van Damme† and Jôel Vandekerckhove†

*Institute of Medical Biochemistry and Danish Center for Human Genome Research, Aarhus University, DK-8000 Aarhus C, Denmark
†Laboratory of Physiological Chemistry, University Ghent, Belgium

INTRODUCTION

The haploid human genome consists of 3 x 10^9 base pairs of DNA distributed in 23 distinct chromosomes. Current estimates indicate that there are about 50,000 to 100,000 genes with perhaps 2,000 to 4,000 genes in each chromosome (Figure 1) (1). Fortunately, only a fraction of the total number of genes is expressed in a distinct cell type at any given time, with perhaps no more than 5,000 different expressed proteins together with their modified variants per cell. Of these, about 70 to 80% may represent household proteins that are shared by all cell types and that are expressed in variable amounts. Assuming that there are at least 250 different cell types in the adult human body (2), each differing from the rest in about 300 to 400 proteins, one ends up with a total number of polypeptides that is reasonably close to the putative number of genes.

To date, only a small proportion of the total set of proteins from any mammalian organism has been identified and in particular, little is known about the protein composition of differentiated cell types. Proteins orchestrate most of the cell functions and it is therefore of importance to identify and functionally characterize as many as possible as well as to reveal those related to disease. Furthermore, given the concerted effort to map and sequence the entire human genome, it is urgent to develop strategies to integrate protein and

J. K. Setlow (ed.), *Genetic Engineering*

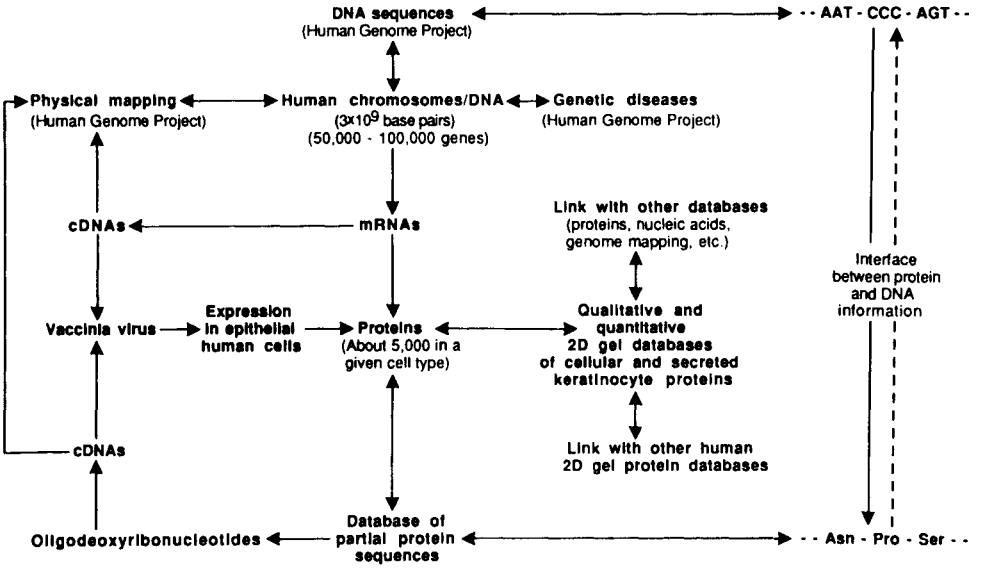


Figure 1. Interface between comprehensive two-dimensional gel protein databases, databases of partial protein sequences and the human genome mapping and sequencing program.

DNA information and to generate comprehensive approaches to the study of the human cell (Figure 1) (3,4).

For the past 15 years, high resolution two-dimensional gel electrophoresis has been the technique of choice to determine the protein composition of a given cell type and for monitoring changes in gene activity through the quantitative and qualitative analysis of the thousands of proteins that orchestrate various cellular functions (3–10 and references therein). The technique which was originally described by O'Farrell (5) separates proteins both in terms of their isoelectric point (pI) and molecular weight. Usually, one chooses a condition of interest and lets the cell reveal the global protein behavioral response as all detected proteins can be analyzed both qualitatively and quantitatively in relation to each other. At present, most available two-dimensional gel techniques (regular gel format) can resolve between 1,000 and 2,000 proteins from a given mammalian cell type, a number that corresponds to about 2 million base pairs of coded DNA. Less abundant proteins can be detected by the analysis of partially purified cellular fractions.

Two-dimensional gel electrophoresis has been widely applied to analysis of cellular protein patterns from bacteria to mammalian cells (3–10, and references therein). In spite of much work, however, information gathered from these studies has not reached the scientific community in its fullness because of lack of standardized gel systems (reagents, equipment, etc.) and the lack of

means for storing and communicating protein information. Only recently, thanks to the development of appropriate computer software, has it been possible to scan gels, assign numbers to individual proteins and store the wealth of information in quantitative and qualitative comprehensive two-dimensional gel protein databases (3,4,9,10,18–33), i.e., those containing information about the various properties (physical, chemical, biological, biochemical, physiological, genetic, immunological, architectural, etc.) of all the proteins that can be detected in a given cell type (Figure 1). Such integrated two-dimensional gel protein databases offer an easy and standardized medium in which to store and communicate protein information and provide a unique framework in which to focus a multidisciplinary approach to the study of the cell. Once a protein is identified in the database, all of the information accumulated can be easily retrieved and made available to the researcher. In the long run, protein databases are expected to foster a wide variety of biological information that may be instrumental to researchers working in many areas of biology—among others, cancer and oncogene studies, differentiation, development, drug development and testing, genetic variation and diagnosis of genetic and clinical diseases (Figure 1).

Systematic two-dimensional gel protein analysis has gained a new dimension with the advent of techniques to microsequence major proteins recorded in the databases (34–53 and references therein). Partial protein sequences can be used to search for protein identity as well as to prepare specific DNA probes for cloning as yet uncharacterized proteins (Figure 1). As these sequences can be stored in the database, they offer a unique opportunity to link protein information with the existing or forthcoming DNA sequence data coming from the human genome project (Figure 1).

A few two-dimensional gel protein databases that are accessible in a computer form have been published *in extenso:* these correspond to the protein-gene database of *Escherichia coli* K-12 developed by Neidhardt and colleagues (18,27,33), the rat REF 52 database established by Garrels and co-workers at Cold Spring Harbor (22,26), the human plasma protein database established by the Andersons (30,54) and a few human databases (transformed amnion cells (19,24,28), normal embryonal lung MRC-5 fibroblasts (21,25), keratinocytes (23,29) and peripheral blood mononuclear cells (19) developed in Aarhus). In addition, there are several smaller cellular databases being established in human (normal human diploid fibroblasts, lymphocytes, leukocytes, leukemic cells) mouse (NIH/3T3 cells, T-lymphocytes), *Aplysia,* yeast (*Saccharomyces cerevisae*), plants (wheat, barley, sorghum) and *Euglena.* Databases of tissue proteins, (brain, whole mouse, liver) and body fluid proteins (plasma proteins, cerebrospinal fluid, urine and milk) are being established in several laboratories. The reader is directed to the review by Celis et al. (9) for details and references concerning these databases. Given space limitations and to keep this review in focus, we will mainly concentrate on the computerized analysis of human cellular two-dimensional gel patterns, and in particular on the steps involved in establishing comprehensive two-dimensional gel databases that will link protein and DNA information.

MAKING AND MANAGING A COMPREHENSIVE TWO-DIMENSIONAL GEL DATABASE OF HUMAN CELLULAR PROTEINS

The first step in making a comprehensive two-dimensional gel protein database is to prepare a synthetic image (digital form of the gel image) of the gel (fluorogram, Coomassie blue or silver-stained gel) to be used as a standard or master reference. This can be done with laser scanners, charge couple device (CCD) array scanners, television cameras, rotating drum scanners and multiwire chambers (17). Computerized analysis systems for spot detection, quantitation, pattern matching and data handling (access and retrieval of information, database making) have been described in the literature (ELSIE (55), GELLAB (15), HERMeS (56), MELANIE (14), QUEST (13), TYCHO (12) and CREAM (57)) and some are available commercially (PDQUEST, Protein Databases Inc., Huntington, N.Y.; KEPLER, Large Scale Biology, Rockville, MD; Visage, BioImage Corporation, Ann Arbor, MI; Gemini, Joyce Loebl, Gateshead; Microscan 1000 Technology Resources Inc., Nashville, TN, MasterScan™, Billerica, MA and CREAM, KEM-EN-TEC, Copenhagen). Unfortunately, most of these systems are incompatible with one another and their advantages and disadvantages have been discussed by Miller (17).

In our workstation in Aarhus, fluorograms are scanned with a Molecular Dynamics laser scanner and the data are analyzed with the use of the PDQUEST II software (Protein Databases Inc.) (13) running on a SPARK station computer 4100 FC-8-P3 from SUN Microsystems, Inc. The scanner measures intensity in the range of 0 to 2.0 absorbance. A typical scan of a 17 cm x 17 cm fluorogram takes about 2 min. Steps in image analysis include: initial smoothing, background subtraction, final smoothing, spot detection and fitting of ideal Guassian distributions to spot centers. Spot intensity is calculated as the integration of a fitted Gaussian. If calibration strips containing individual segments of known amount of radioactivity are used, it is possible to merge multiple exposures of the sample image into a single data image of greater dynamic range. Once the synthetic image is created it can be stored on disk and displayed directly on the monitor. Functions that can be used to edit the images include: cancel (for example to erase scratches that may have been interpreted as spots by the computer; cancel streaks or low dpm spots), combine (sometimes a spot may be resolved into several closely packed spots), restore, uncombine and add spot to the gel. The editing process is time consuming—about 1 to 1 1/2 day per image. Figure 2A shows a standard synthetic image (IEF, isoelectric focusing) of a fluorogram of [^{35}S]-methionine labelled cellular proteins from human keratinocytes (master database) (29). Images can be displayed either in black and white (resembling the original fluorograms) or in color depending on the need. As shown in Figure 2B, each polypeptide is assigned a number by the computer, which facilitates the entry and retrieval of qualitative and quantitative information for any given spot in the gel (29). The standard image can be matched automatically by the computer to other standard or reference gels provided a few landmark spots

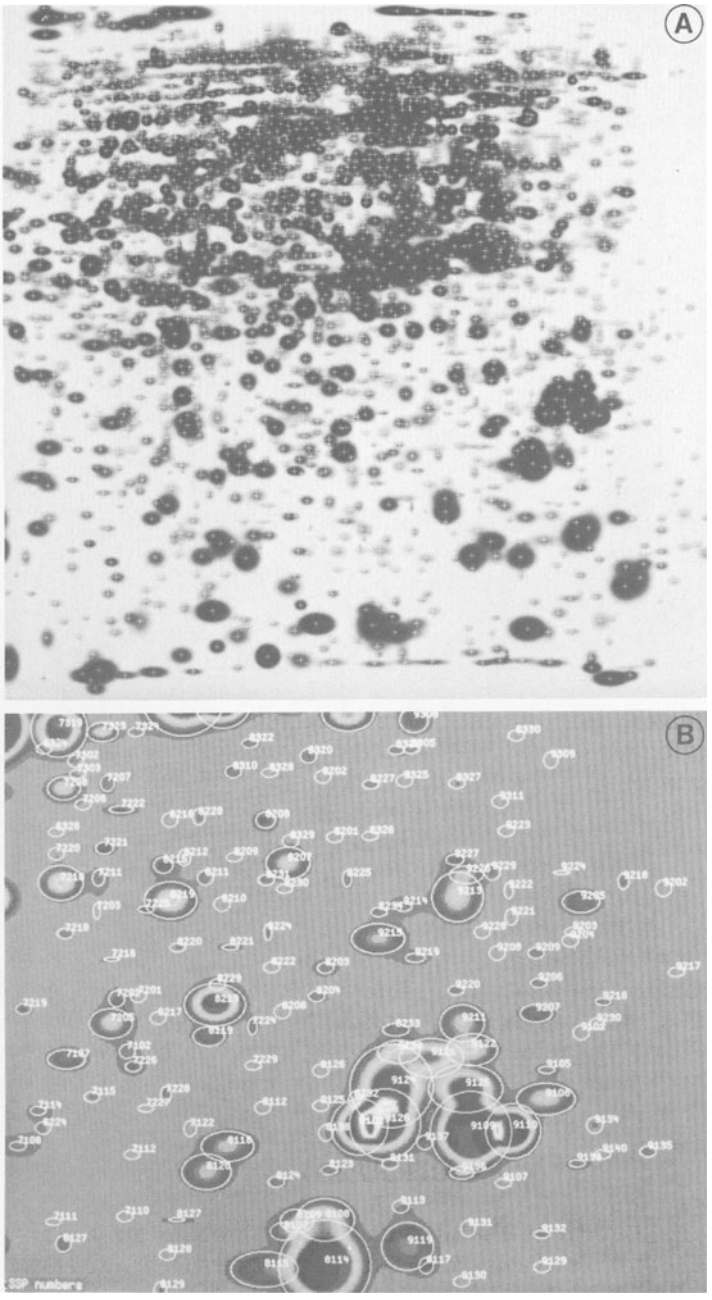


Figure 2 A) Synthetic image (master image) of a fraction of an IEF fluorogram of [^{35}S]-methionine-labelled proteins from human keratinocyte proteins. B) Image showing numbers assigned by the computer to each spot.

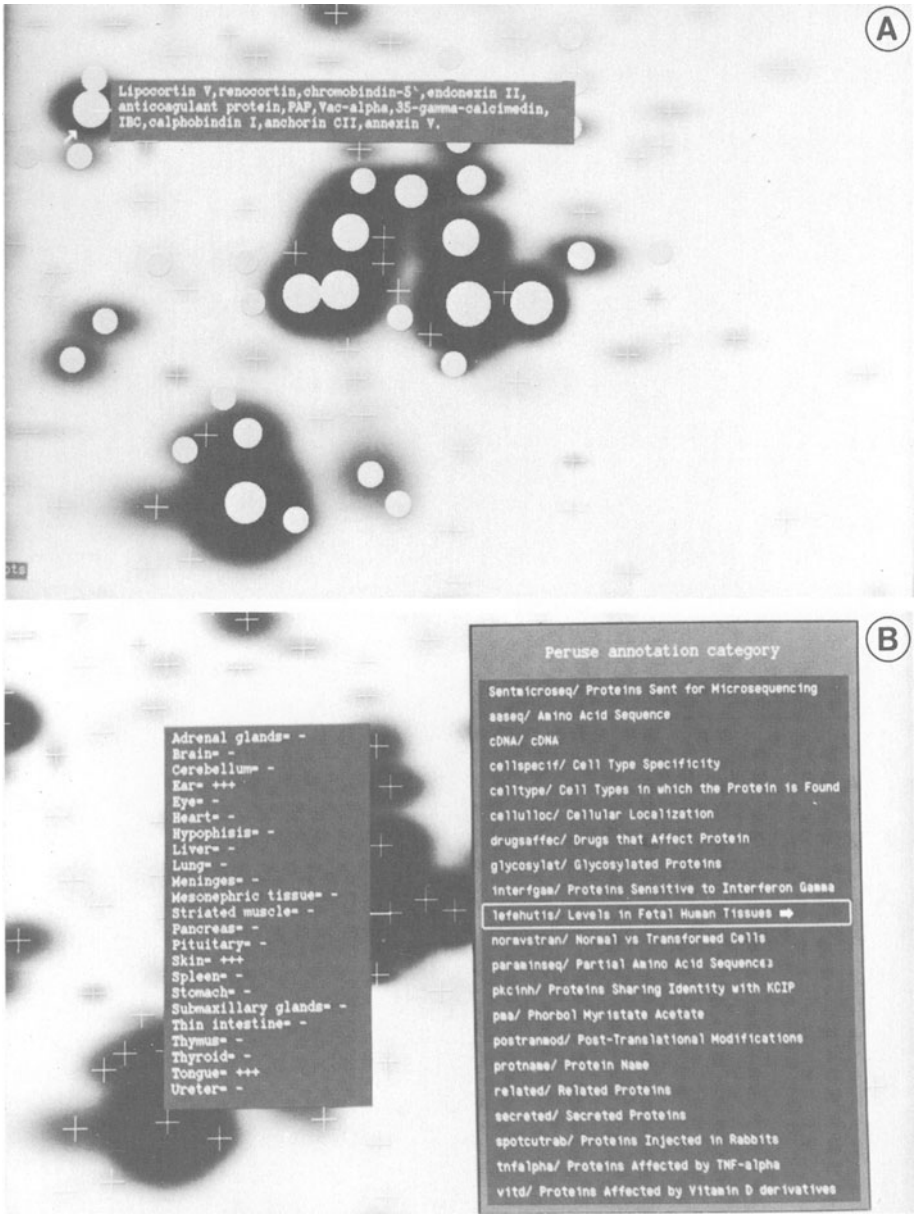


Figure 3 A) Synthetic image of a fraction of an IEF fluorogram showing proteins (flagged with white) that contain information under the entry *protein name.* Only the information contained in one of the spots is displayed. About 600 proteins contain information in this category. B) As A) but using the function *peruse annotation for spot* to inquire about categories and information available for the epithelial marker stratifin.

are given manually as reference to initiate the process. It should be stressed that proteins are matched according to their gel position and, therefore, additional ways to verify their relatedness are needed before one can take full advantage of these data.

Once a standard map of a given protein sample is made, one can enter qualitative annotations to establish a reference database. Our current master two-dimensional gel database of human keratinocyte cellular proteins (update 1992; this database is updated yearly in Electrophoresis) lists 2,980 polypeptides having pI's ranging from 4 to 13 and molecular weights between 8.5 and 230 kD. The most abundant protein in the database corresponds to total actin (about 90 million molecules per cell) while the less abundant of the recorded polypeptides are present in around 5,000 molecules per cell. Some annotation categories that we are using to establish the master keratinocyte database include: 1) protein identification (co-migration with purified proteins, two-dimensional gel immunoblotting, microsequencing, cDNA expression in the vaccinia virus expression system); 2) amounts (total amounts and levels of synthesis); 3) subcellular localization (nuclear, cytoskeletal, membrane, membrane receptors, specific organelles, etc.); 4) antibody against protein; 5) post-translational modifications (phosphorylation, glycosylation, myristoylation, etc.); 6) microsequencing; 7) cell cycle specificity (specific variations in levels of synthesis and amount); 8) regulatory behavior (effect of hormones, growth factors, cytokines, heat shock, etc.); 9) rate of synthesis in normal and transformed cells (proliferation sensitive proteins, cell cycle specific proteins, differentiation markers, oncogenes, components of the pathways that control cell proliferation); 10) function (mainly from co-migration with proteins of known function); 11) sets of proteins that are coordinately regulated (hierarchy of controls, differential gene expression in various stages of differentiation, etc.); 12) cDNAs (cDNAs cloned in the laboratory); 13) proteins that are specific to a given disease (for example systematic comparison of protein patterns from normal and psoriatic keratinocytes); 14) expression and exploitation of transfected cDNAs; 15) pathways (signal transduction, metabolic, others); 16) gene localization (genetic and physical) and 17) effect of microinjected antibody on patterns of protein synthesis.

Information entered for any spot in a given annotation category can be easily retrieved by asking the computer to display the information on the screen. For example, Figure 3A shows a synthetic image of an IEF gel (master keratinocyte database) displaying the information contained under the entry *protein name.* All the spots flagged with white in Figure 3A contain information in this category. By using the function *show annotation for spot* it is possible to display the annotation for a particular protein. Alternatively, one can use the function *peruse annotations for spot* directly to ask the computer to list all the entries available for a particular protein (in this case stratifin). By clicking the mouse in a given entry (presence in fetal human tissues) it is possible to take a quick look at the information in that particular entry (Figure 3B). Table 1 lists entries available for the epithelial marker stratifin.

One of the short-term goals we have set in building the master database of human keratinocyte proteins (29) has been the identification of as many proteins as possible. This is currently being done by one or a combination of the following procedures: 1) co-migration with known proteins, 2) 2-D gel immunoblotting with the use of specific antibodies, 3) microsequencing of Coomassie brilliant blue stained human proteins recovered from dried 2-D gels and 4) expression of known cDNAs in the vaccinia virus expression system. So far, we have received about 900 antibodies from laboratories all over the world and these are being systematically tested by 2-D gel immunoblotting for antigen determination. Similarly, purified proteins and organelles provided by several laboratories have greatly aided the identification of proteins. We routinely request antibodies and protein samples and promise the donors to make available all of the information we may have accumulated on that particular protein. In general, the process of data collection and dissemination is laborious and requires substantial manpower. To date, about 600 proteins (~20% of the total number of proteins recorded) have been identified in the keratinocyte database.

As mentioned earlier, one distinct advantage of two-dimensional gel electrophoresis is the possibility of studying quantitative variations in cellular protein patterns that may lead to the identification of groups of proteins that are expressed coordinately during a given biological process. Quantitation, however, is not an easy task as reflected by the lack of published data on global cellular protein patterns. We believe that this is partly due to difficulties in obtaining sets of gels that are suitable for computer analysis (streaking, material remaining at the origin, etc.) as well as to limitations (laborious editing time, need of calibration strips to merge images, limited dynamic range, etc.) in the computer analysis systems available at the moment. Perhaps the most advanced quantitative studies published so far with computer analysis have been carried out by Garrels and co-workers (22,26). In particular, these investigators have established a quantitative rat protein database (22,26) designed to study growth control (proliferation, growth inhibitors and stimulation) and transformation in well-defined groups of cell lines obtained by transformation of rat REF52 cells with SV40, adenovirus and the Kirsten murine sarcoma virus. These studies have revealed clusters of proteins induced or repressed during growth to confluence as well as groups of transformation-sensitive proteins that respond in a differential fashion to transformation by DNA and RNA viruses. A most interesting feature of this quantitative database is the discovery of a group of co-regulated proteins that show similar expression patterns as the cell cycle-regulated DNA replication protein known as proliferating cell nuclear antigen (PCNA)/cyclin (58).

In our human databases, most quantitations have been carried out by estimating the radioactivity contained in the polypeptides by direct counting of the gel pieces in a scintillation counter (24,25). Up to 700 proteins can be cut out through appropriate exposed films in a period of time comparable to that required for editing a synthetic image. Manual quantitation of this large number of spots is difficult, however, without the assistance of a master

Table 1
Some Entries for Stratifin in the Human Keratinocyte Two-dimensional Gel Protein Database

Entries for Stratifin (IEF SSP 9109)	Information Entered
1. Protein name	Stratifin
2. Apparent molecular weight (Mr)	30.0 kD
3. Isoelectric point (pI)	4.4
4. Method (or methods) of identification	Microsequencing, cDNA cloning and expression
5. Antibody against protein	Polyclonal (rabbit) J.E. Celis and B. Basse, Aarhus
6. Cellular localization	Cellular and secreted
7. Partial amino acid sequence	YEDMAAF (19-25), NLLSVAYK (42-49), VFYLK (118-122), YLAEVATGDD (130-139) LGLALNFSVFXY (170-181), XYEIANSPEE(A)I (180-191), DNLTL (225-229), (T/A)ADNAGEEG (231-239)
8. cDNA sequence	Known. H. Leffers et al. (60). (Mr = 27,773 D, pI = 4.5 from translated sequence.)
9. Levels in fetal human tissues	Adrenal glands = -; brain = -; cerebellum = -; ear = +++; eye = -; heart = -; hypophysis = -; liver = -; lung = -; meninges = -; mesonephric tissue = -; striated muscle = -; pancreas = -; skin = +++; spleen = -; stomach = -; submandibular gland = -; small intestine = -; thymus = -; thyroid gland = -; tongue = +++; ureter = -.
10. Levels in quiescent, proliferating and transformed MRC-5 fibroblasts	Q (quiescent) = 1.3; P (proliferating) = 1.0; T (SV40 transformed) = 0.1
11. Cell specificity	Only in cultured epithelial cells
12. Modification	Glycosylated
13. Phorbol esters	Down regulated

reference image and a numbering system that can be used to identify the spots. Using this approach, we have recorded quantitative changes in the relative abundance of 592 [^{35}S]-methionine-labelled proteins synthesized by quiescent, proliferating, and SV40-transformed human embryonic lung MRC-5 fibroblasts (25). At present, we are investigating the possibility of using the phosphoimager to obtain quantitative data. Our studies as well as those of Garrels and co-workers (22,26) may in the long run help define patterns of gene expression that are characteristic of the transformed state.

MICROSEQUENCING HAS ADDED A NEW DIMENSION TO COMPREHENSIVE TWO-DIMENSIONAL GEL DATABASES: A DIRECT LINK BETWEEN PROTEINS AND GENES

The development of highly-sensitive amino acid gas-phase or liquid-phase sequenators (24), together with the establishment of efficient protein and peptide sample preparation methods, has opened the possibility to perform a

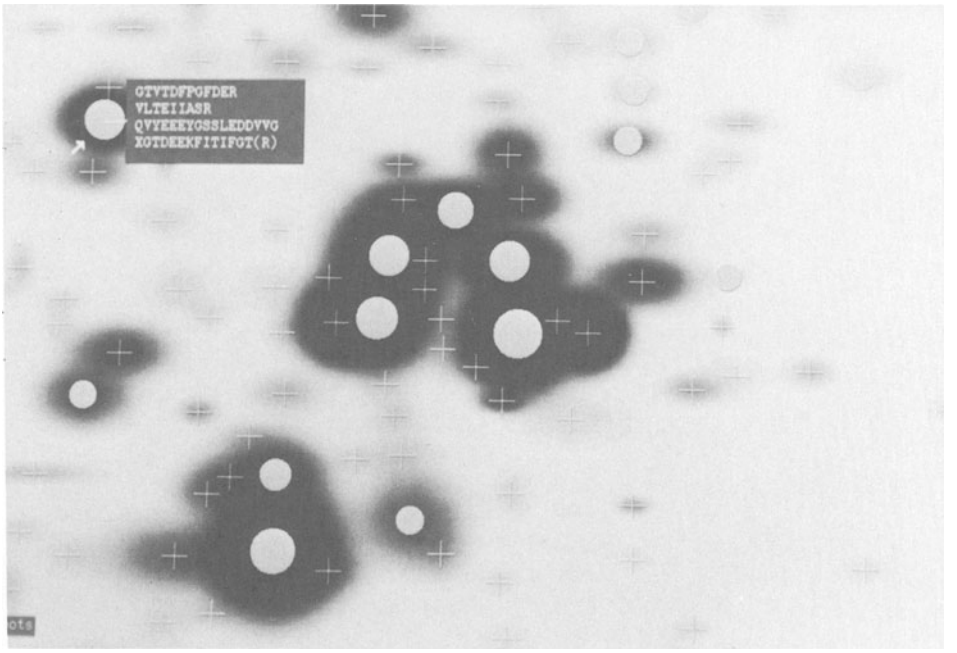


Figure 4. Synthetic image of a fraction of an IEF fluorogram showing polypeptides that contain information under the category of *partial amino acid sequences.* Only the information contained in one of the spots is displayed. About 250 proteins contain information in this category.

systematic sequence analysis of proteins resolved by two-dimensional gel electrophoresis. Indeed, generated pieces of protein sequences can be used to search for protein identity (comparison with available sequences stored in databanks), as well as for preparing specific DNA probes for cloning of as yet uncharacterized proteins (Figure 1). In addition, partial protein sequences can be stored in two-dimensional gel databases (Figure 4; all the proteins flagged with white contain information in this category) and offer a unique link between proteins and genes (Figure 1).

In the early 1970s, gel electrophoresis was used to purify proteins for sequencing purposes (reviewed by Weber and Osborn in (35)). Proteins were recovered by diffusion and sequenced by the manual dansyl-Edman degradation at the nanomole level. This technique was further refined by using electro-elution to recover proteins and by miniaturizing the system (36). This method has been used extensively, but showed increasing drawbacks (low yields, protein samples contaminated by free amino acids, and NH_2-terminal blocking) as the amounts of handled protein gradually became smaller (e.g., at the picomole level).

The introduction of protein-electroblotting procedures (37–42) and chemically inert membranes made it possible to sequence the immobilized proteins directly without additional manipulations, generally yielding NH_2-terminal sequences containing 10 to 40 residues. A major difficulty encountered in this procedure is the occurrence of frequent artifactual blockage of the proteins. In addition to this primarily technical problem, many proteins are blocked *in vivo* by acylation or by a pyrrolidone carboxylic acid cap.

The problem of partial or complete NH_2-terminal blockage can be circumvented by generating internal amino acid sequences. This is achieved by fragmenting the protein present in the gel (gel *in situ* cleavage), or by cleaving it while bound to the membrane (membrane *in situ* cleavage) (43–45). This method has been described for Ponceau red-stained proteins on nitrocellulose blots (44), for Amido-black-stained Immobilon-bound proteins, and for fluorescamine-detected proteins on glass-fiber membranes (45). The proteases used (trypsin, chymotrypsin, V8 or pepsin) cleave at multiple sites, generating small peptides which elute from the blot into the digestion buffer from which they are purified by reverse-phase high performance liquid chromatography (HPLC) before being sequenced individually. As membrane-immobilized proteins are not homogeneously digested, but rather show protease sensitivity next to resistant regions, the number of peptides generated is much lower than expected from the number of potential cleavage sties. Consequently, HPLC peptide chromatograms are less complex and most peptides can be recovered in pure form.

As only limited amounts of a protein mixture can be loaded on a two-dimensional gel, proteins of interest are often obtained in yields insufficient for the currently available sequencing technology. More material can be obtained by enriching for a certain subcellular fraction (purified cell organelles) or by exploiting affinity (dyes, metals, drugs, etc.) or hydrophobic properties of proteins before gel analysis. All of the sequencing results accumulated so far in the human protein databases (24,29,53) (an example is shown in Figure 4)

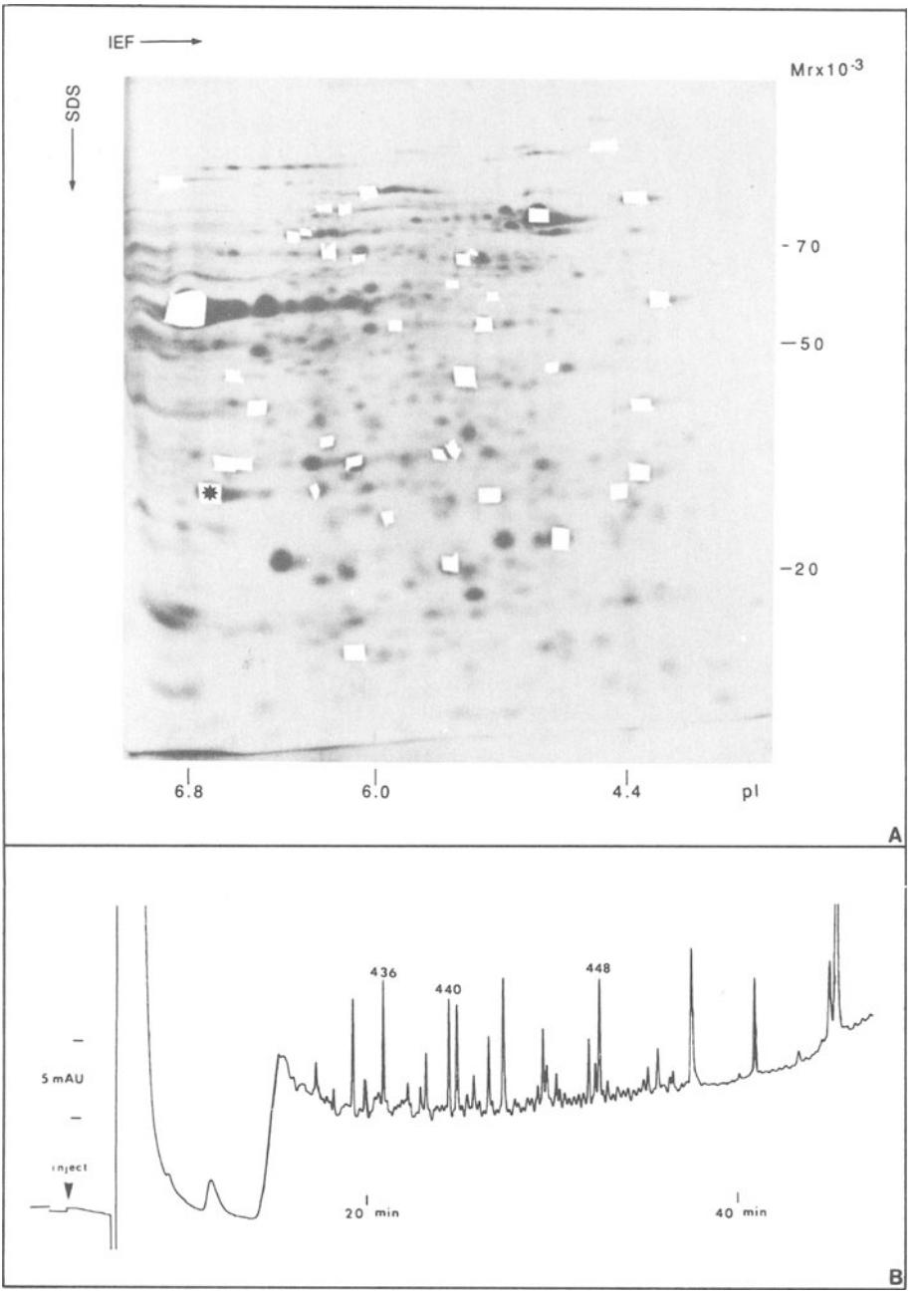
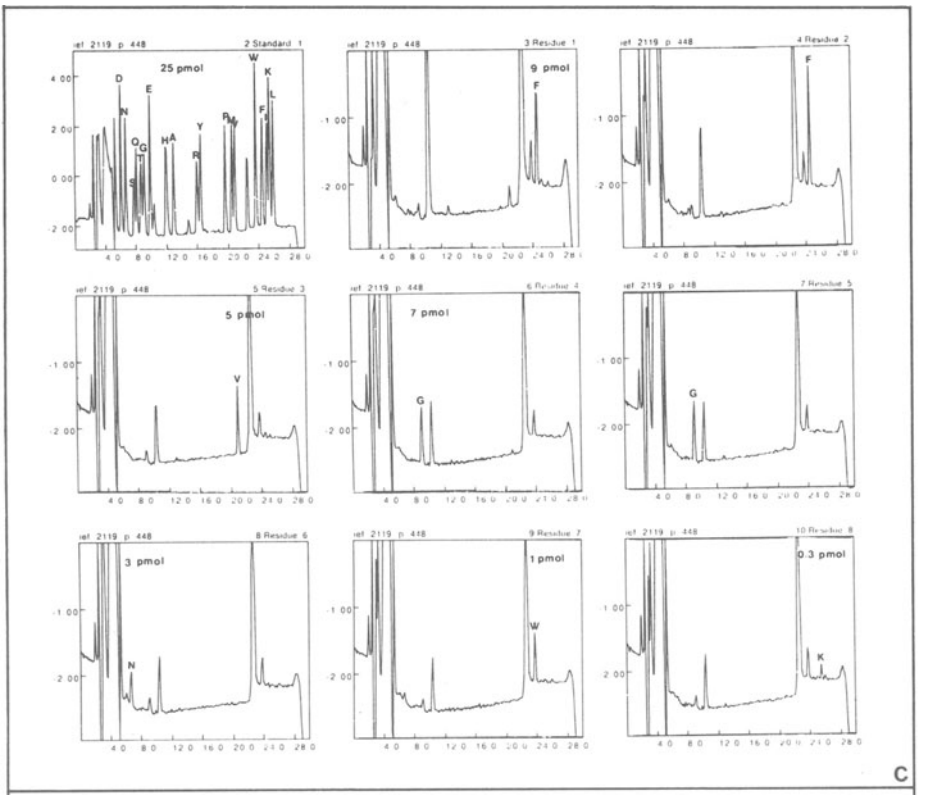


Figure 5. The various steps involved in microsequencing are illustrated by the analysis of protein IEF SSP 2119 (24). (A) Representative Coomassie brilliant blue stained gel of partially purified Molt-4 proteins (about 1 mg of total protein was loaded) used to cut protein IEF SSP 2119 (indicated with an asterisk). Spots from 6 gels were used in this particular analysis. (B) HPLC chromatogram of peptides generated by *in situ* digestion of electroblotted protein IEF SSP 2119. The column was a C18 normal-bore. (C) Phenylthio-




```
ID   TPIS$HUMAN      STANDARD;       PRT;   248 AA.

DE   TRIOSEPHOSPHATE ISOMERASE (EC 5.3.1.1) (GENE NAME: TPI).
OS   HUMAN (HOMO SAPIENS).

  Sw:Tpis$Human  Length: 248
                    448
    1  APSRKFFVGG NWKMNGRKQS LGELIGTLNA AKVPADTEVV CAPPTAYIDF
                 440a
   51  ARQKLDPKIA VAAQNCYKVT NGAFTGEISP GMIKDCGATW VVLGHSERRH
                                                  440b
  101  VFGESDELIG QKVAHALAEG LGVIACIGEK LDEREAGITE KVVFEQTKVI
                          436
  151  ADNVKDWSKV VLAYEPVWAI GTGKTATPQQ AQEVHEKLRG WLKSNVSDAV

  201  AQSTRIIYGG SVTGATCKEL ASQPDVDGFL VGGASLKPEF VDIINAKQ


  PEPTIDESORT of: Tpis$Human.Sw
       Summary for whole sequence:

  Molecular weight =   26538.14
  Isoelectric point =  6.89
```

D

hydantoin (PTH) chromatograms from amino acid sequencing of the tryptic peptide 448. Data were collected on-line with the 610A data system. (D) Amino acid sequences from three tryptic peptides were used for homology searches that gave total homology to human triosephosphate isomerase. The molecular weight and pI calculated from the published sequence matched completely the gel coordinates recorded in the master AMA two-dimensional gel protein database (24).

have been obtained from the analysis of protein spots collected from two-dimensional gels that had been stained with Coomassie brilliant blue according to standard procedures and dried for storage. Proteins are recovered from the collected gel pieces by elution-concentration gel electrophoresis. Details of this technique have been reported previously (52,53) and a brief outline is given below (see also Figure 5).

Combined gel pieces are allowed to swell in gel sample buffer (a total volume of 1.5 ml). The gel pieces combined with the supernatant are then loaded into the slot of an elution-concentration gel. In this way the protein is efficiently eluted from the gel pieces and concentrated from a large volume into a narrow spot as a result of both a vertical stacking and horizontal contraction of the protein band. The highly concentrated (about 5 mm^2) protein spot is then electroblotted on polyvinyldiene difluoride (PVDF)-membranes, stained with Amido black and *in situ* digested with trypsin. The peptides generated during digestion elute from the membrane and are separated by normal or narrow bore reverse-phase HPLC and collected individually for sequence analysis. In our hands, this approach is routinely applied to gel-purified proteins available in few µgs. The various steps involved in microsequencing are illustrated in Figure 5 (analysis of protein IEF SSP 2119).

Using this and previous procedures (47,49,52,53), we have so far microsequenced 250 protein spots collected from two-dimensional gels (24,47,49,52,53 and unpublished observations), a number we expect to increase by at least 100 per year. Of the microsequenced proteins about 50% corresponded to unknown proteins.

GENERAL CONSIDERATIONS AND PERSPECTIVES

One of the major advantages in developing computer-accessible human 2-D gel databases in which most of the known proteins are identified is the wealth of new proteins that will become amenable to experimentation both at the biochemical and molecular biology level. Accordingly, a high-priority goal in our program has been to establish a database of partial sequences of unknown proteins that may be interfaced with the forthcoming DNA sequence information from the Human Genome Project (Figure 1). Besides aiding the identification of genes, these sequences can be used to prepare oligodeoxyribonucleotides which in turn are used to clone the corresponding cDNAs (Figure 1). Complete cDNAs can then be sequenced and used for physical mapping in collaborative studies. Many proteins, however, are members of families, and therefore it is necessary to determine which protein is actually coded for by a given cDNA. This has been achieved by expressing the cDNA in human epithelial amnion (AMA) cells with the use of a vaccinia virus vector that turns off the host cells' protein synthesis while producing large quantities of the proteins coded in the virus genome (59–61 and references therein) (Figure 6). The protein coded by the cDNA is then identified by superimposing the autoradiogram of the proteins synthesized by the recombinant virus-infected cells with the silver-stained gel showing the "background" of unlabelled cellular

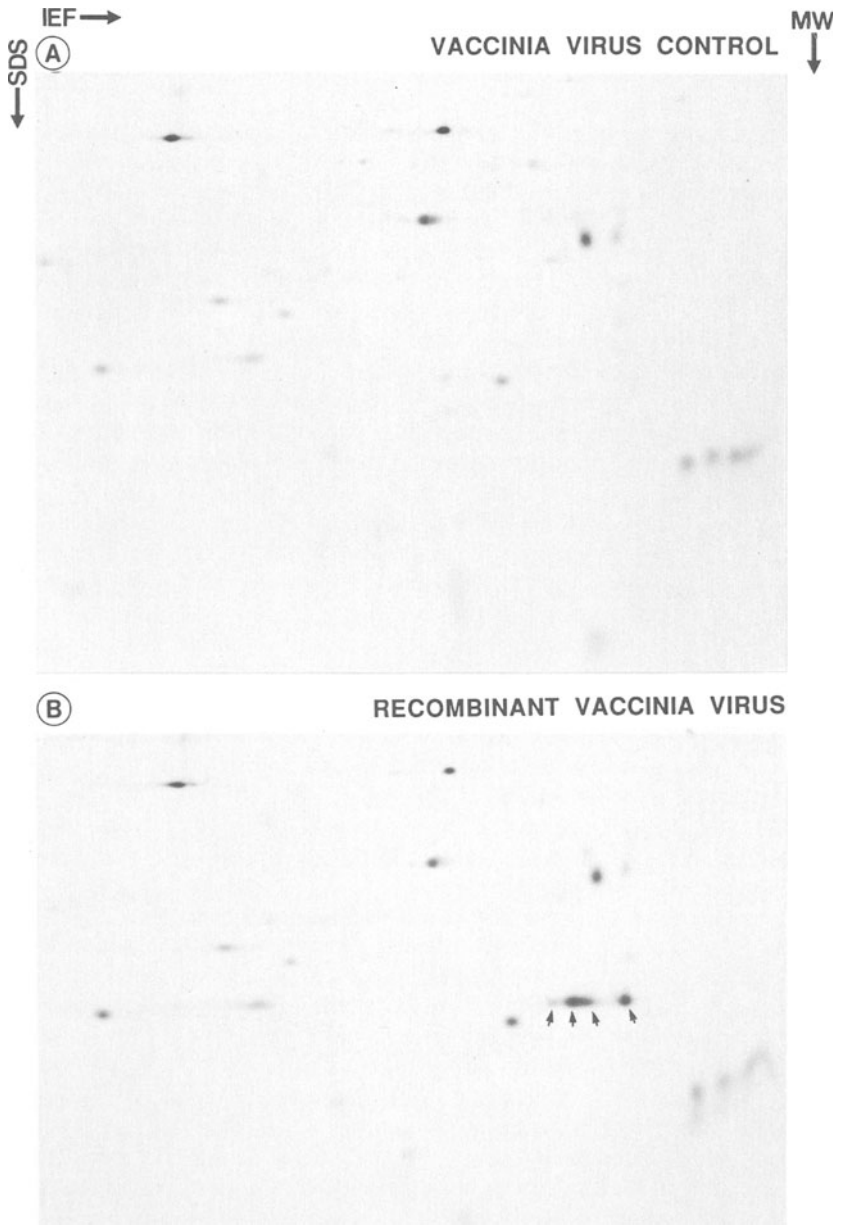


Figure 6. IEF autoradiograms of [^{35}S]-methionine-labelled proteins synthesized by AMA cells infected with (A) control vaccinia virus and (B) vaccinia virus carrying the 8106 clone (60).

proteins. The method also allows us to identify other cloned proteins in the protein databases by either asking researchers to send their cDNA clones to us or by cloning the cDNAs ourselves using the PCR method. So far we have cloned and sequenced about 15 specific cDNAs.

An alternative approach to increase the number of cDNAs ("non-specific" cDNAs, Figure 1) that will be available for physical mapping is currently being pursued by H. Leffers. The procedure is as follows: we pick cDNA clones at random from *Not* I-poly [dT] primed libraries, sequence both ends of the insert, analyze the insert size by gel electrophoresis, check if they are full length by Northern blotting, recombine the full-length cDNAs into vaccinia virus and determine which polypeptide they code for using two-dimensional gel electrophoresis. The main obstacle in the random approach, however, is the large amount of DNA sequencing required. We recently started a collaborative project with W. Ansorge's group at the European Molecular Biology Laboratory (EMBL; sponsored by the European Community Human Genome Analysis Program) to sequence some of the cDNAs using automated DNA sequencing machines.

SUMMARY

Analysis of cellular protein patterns by computer-aided two-dimensional gel electrophoresis together with recent advances in protein sequence analysis and expression systems have made possible the establishment of comprehensive two-dimensional gel protein databases that may link protein and DNA mapping and sequence information and that offer an integrated approach to the study of gene expression. With the integrated approach offered by two-dimensional gel protein databases it is now possible to reveal phenotype-specific protein(s), to microsequence them, to search for homology with previous identified proteins, to clone the cDNAs, to assign partial protein sequences to genes for which the full DNA sequence and the chromosome location are known, and to study the regulatory properties and function of groups of proteins that are coordinately expressed in a given biological process. Comprehensive two-dimensional gel protein databases will provide an integrated picture of the expression levels and properties of the thousands of protein components of organelles, pathways, and cytoskeletal systems, both under physiological and abnormal conditions, and are expected to lead to the identification of new regulatory networks.

So far, about 20% (600 out of 2,980) of the total number of proteins recorded in the human keratinocyte protein database have been identified and we are actively gathering qualitative and quantitative biological data on all resolved proteins. Given the current improvements on microsequencing as well as the availability of specific antibodies, it seems feasible to expect that most known keratinocyte proteins will be identified in the very near future. This feast will reveal a wealth of new proteins that will become amenable to experimentation both at the biochemical and molecular biology level.

REFERENCES

1 Human Genome II Conference, San Diego, 1990.
2 Geneser, F. (1986) Textbook of Histology, Munksgaard, Copenhagen.
3 Special issue: Two-dimensional Gel Protein Databases (1990) Electrophoresis 11, 989–1168.
4 Celis, J.E., Rasmussen, H.H., Leffers, H., Madsen, P., Honoré, B., Gesser, B., Dejgaard, K. and Vandekerckhove, J. (1991) FASEB J. 5, 2200–2208.
5 O'Farrell, P.H. (1975) J. Biol. Chem. 250, 4007–4021.
6 O'Farrell, P.Z., Goodman, H.M. and O'Farrell, P.H. (1977) Cell 12, 1133–1142.
7 Special Issue: Two-dimensional Gel Electrophoresis (1982) Clin. Chem. 28, 737–1092.
8 Celis, J.E. and Bravo, R. (eds.) (1984) Two-Dimensional Gel Electrophoresis of Proteins: Methods and Applications. Academic Press, New York, NY.
9 Celis, J.E., Madsen, P., Gesser, B., Kwee, S., Nielsen, H.V., Rasmussen, H.H., Honoré, B., Leffers, H., Ratz, G.P., Basse, B., Lauridsen, J.B. and Celis, A. (1989) In Advances in Electrophoresis 3 (Chrambach, C., Dunn, M.J. and Radola, B.J., eds.) VCH, Weinheim.
10 Special Issue: Two-dimensional Gel Protein Databases (Celis, J.E., ed.) (1990) Electrophoresis 11, 989–1166.
11 Garrels, J.I. (1983) Methods Enzymol. 100, 411–423.
12 Anderson, N.L., Hofmann, J.P., Gemmell, A. and Taylor, J. (1984) Clin. Chem. 30, 2031–2036.
13 Garrels, J.I., Farrar, J.T. and Burwell, C.B. (1984) in Two-dimensional Gel Electrophoresis of Proteins. Methods and Applications (Celis, J.E. and Bravo, R., eds.) pp. 37–91, Academic Press, New York, NY.
14 Vincens, P. and Tarroux, P. (1988) Internat. J. Biochem. 20, 499–509.
15 Appel, R., Hochstrasser, D., Roch, C., Funk, M., Muller, A.F. and Pellegrini, C. (1988) Electrophoresis 9, 136–142.
16 Lemkin, P.F. and Lester, E.P. (1989) Electrophoresis 10, 122–139.
17 Miller, M.J. (1989) in Advances in Electrophoresis 3 (Chrambach, C., Dunn, M.J. and Radola, B.J., eds.) pp. 182–217, VCH, Weinheim.
18 Philips, T.D., Vaughn, V., Bloch, P.L. and Neidhardt, F.C. (1987) in *Escherichia coli* and *Salmonella typhimurium,* Cellular and Molecular Biology, Gene-Protein Index of *Escherichia coli* K-12, 2nd ed. (Neidhardt, F.C., Ingraham, J.I., Low, K.B., Magasanik, B., Schaechter, M. and Umbarger, H.E., eds.) pp. 919–966, American Society for Microbiology, Washington, DC.
19 Celis, J.E., Ratz, G.P., Celis, A., Madsen, P., Gesser, B., Kwee, S., Nielsen, H.V., Yde, H., Lauridsen, J.B. and Basse, B. (1988) Leukemia 2, 561–601.
20 Special Issue: Protein Databases in Two-dimensional Electrophoresis (Celis, J.E., ed.) (1989) Electrophoresis 10, 73–164.
21 Celis, J.E., Ratz, G.P., Madsen, P., Gesser, B., Lauridsen, J.B., Brogaard Hansen, K.P., Kwee, S., Rasmussen, H.H., Nielsen, H.V., Crüger, D.,

Basse, B., Leffers, H., Honoré, B., MΦller, O. and Celis, A. (1989) Electrophoresis 10, 76–115.
22 Garrels, J.I. and Franza, B.R. (1989) J. Biol. Chem. 264, 5283–5298.
23 Celis, J.E., Crüger, D., Kiil, J., Dejgaard, K., Lauridsen, J.B., Ratz, G.P., Basse, B., Celis, A., Rasmussen, H.H., Bauw, G. and Vandekerckhove, J. (1990) Electrophoresis 11, 242–254.
24 Celis, J.E., Gesser, B., Rasmussen, H.H., Madsen, P., Leffers, H., Dejgaard, K., Honoré, B., Olsen, E., Ratz, G., Lauridsen, J.B., Basse, B., Mouritzen, S., Hellerup, M., Andersen, A., Walbum, E., Celis, A., Bauw, G., Puype, M., Van Damme, J. and Vandekerckhove, J. (1990) Electrophoresis 11, 989–1071.
25 Celis, J.E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honoré, B., Rasmussen, H.H., Olsen, E., Lauridsen, J.B., Ratz, G., Mouritzen, S., Basse, B., Hellerup, M., Celis, A., Puype, M., Van Damme, J. and Vandekerckhove, J. (1990) Electrophoresis 11, 1072–1113.
26 Garrels, J.I., Franza, B.R., Chang, C. and Latter, G. (1990) Electrophoresis 11, 1114–1130.
27 VanBogelen, R.A., Hutton, M.E. and Neidhardt, F.C. (1990) Electrophoresis 11, 1131–1166.
28 Celi, J.E., Leffers, H., Rasmussen, H.H., Madsen, P., Honoré, B., Gesser, B., Dejgaard, K., Olsen, E., Ratz, G.P., Lauridsen, J.B., Basse, B., Andersen, A.H., Walbum, E., Brandstrup, B., Celis, A., Puype, M., Van Damme, J. and Vandekerckhove, J. (1991) Electrophoresis 12, 765–801.
29 Celis, J.E., Madsen, P., Rasmussen, H.H., Leffers, H., Honoré, B., Gesser, B., Dejgaard, K., Olsen, E., Magnusson, N., Kiil, J., Celis, A., Lauridsen, J.B., Basse, B., Ratz, G.P., Andersen, A.H., Walbum, E., Brandstrup, B., Pedersen, P.S., Brandt, N.J., Puype, M., Van Damme, J. and Vandekerckhove, J. (1991) Electrophoresis 12, 802–872.
30 Anderson, N.L. and Anderson, N.G. (1991) Electrophoresis 12, 883–906.
31 Anderson, N.L., Esquer-Blasco, R., Hofmann, J.-P. and Anderson, N.G. (1991) Electrophoresis 12, 907–930.
32 Wirth, P.J., Luo, L., Fujimoto, Y., Bisgaard, H.C. and Olson, A.D. (1991) Electrophoresis 12, 931–954.
33 VanBogelen, R.A. and Neidhardt, F.C. (1991) Electrophoresis 12, 955–994.
34 Hewick, R.M., Hunkapiller, M.W., Hood, L.E. and Dreyer, W.J. (1981) J. Biol. Chem. 256, 7990–7997.
35 Weber, K., and Osborn, M. (1985) in The Proteins and Sodium Dodecyl Sulfate, Molecular Weight Determination on Polyacrylamide Gels and Related Procedures. (Neurath, H. et al., eds.) Vol. 1, pp. 179–223, Academic Press, New York, NY.
36 Hunkapiller, M.W., Lujan, E., Ostrander, F. and Hood, L.E. (1983) Methods Enzymol. 91, 227–236.
37 Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J. and Van Montagu, M. (1985) Eur. J. Biochem. 152, 9–19.
38 Aebersold, R.H., Teplow, D.B., Hood, L.E. and Kent, S.B.H. (1986) J. Biol. Chem. 261, 4229–4238.

39 Bauw, G., De Loose, M., Inzé, D., Van Montagu, M. and Vandekerckhove, J. (1987) Proc. Nat. Acad. Sci. U.S.A. 84, 4806–4810.
40 Matsudaira, P. (1987) J. Biol. Chem. 262, 10035–10038.
41 Eckerskorn, C., Mewes, W., Goretzki, H. and Lottspeich, F. (1988) Eur. J. Biochem. 176, 509–519.
42 Moose, M., Jr., Nguyen, N.Y. and Liu, T.-Y. (1988) J. Biol. Chem. 263, 6005–6008.
43 Kennedy, T.E., Gawinowicz, M.A., Barzilai, A., Kandel, E.R. and Sweatt, J.D. (1988) Proc. Nat. Acad. Sci. U.S.A. 85, 7008–7012.
44 Aebersold, R.H., Leavitt, J., Saavedra, R.A., Hood, L.E. and Kent, S.B.H. (1987) Proc. Nat. Acad. Sci. U.S.A. 84, 6970–6974.
45 Bauw, G., Van den Bulcke, M., Van Damme, J., Puype, M., Van Montagu, M. and Vandekerckhove, J. (1988) J. Prot. Chem. 7, 194–196.
46 Celis, J.E., Ratz, G.P., Madsen, P., Gesser, B., Lauridsen, J.B., Kwee, S., Rasmussen, H.H., Nielsen, H.V., Crüger, D., Basse, B., Leffers, H., Honoré, B., Mϕller, O., Celis, A., Vandekerckhove, J., Bauw, G., Van Damme, J., Puype, M. and Van den Bulcke, M. (1989) FEBS Lett. 244, 247–254.
47 Bauw, G., Van Damme, J., Puype, M., Vandekerckhove, J., Gesser, B., Ratz, G.P., Lauridsen, J.B. and Celis, J.E. (1989) Proc. Nat. Acad. Sci. U.S.A. 86, 7701–7705.
48 Aebersold, R. and Leavitt, J. (1990) Electrophoresis 11, 517–527.
49 Bauw, G., Rasmussen, H.H., Van den Bulcke, M., Van Damme, J., Puype, M., Gesser, B., Celis, J.E. and Vandekerckhove, J. (1990) Electrophoresis 11, 528–536.
50 Tempst, P., Link, A.J., Riviere, L.R., Fleming, M. and Elicone, C. (1990) Electrophoresis 11, 537–553.
51 Eckerskorn, C. and Lottspeich, F. (1990) Electrophoresis 11, 554–561.
52 Rasmussen, H.H., Van Damme, J., Bauw, G., Puype, M., Gesser, B., Celis, J.E. and Vandekerckhove, J. (1991) in Methods in Protein Sequence Analysis (Jörnvall, H., Höög, J.-O. and Gustavsson, A.-M., eds.) pp. 103–114, Birkhäuser Verlag, Basel.
53 Rasmussen, H.H., Van Damme, J., Puype, M., Gesser, B., Celis, J.E. and Vandekerckhove, J. (1991) Electrophoresis 12, 873–882.
54 Anderson, N.G. and Anderson, N.L. (1979) Behring. Inst. Mitt. 63, 169–210.
55 Olson, A.D. and Miller, M.J. (1988) Anal. Biochem. 169, 49–70.
56 Vincens, P., Paris, N., Pujol, J.L., Gaboriaud, C., Rabilloud, T., Pennetier, J., Matherat, P. and Tarroux, P. (1986) Electrophoresis 7, 347–356.
57 Hagerup, M., Conradsen, K., Bϕgh-Hansen, T.C., Bouchelouche, P. and Schafer-Nielsen, C. (1988) Electrophoresis '88, Sixth Meeting of the International Electrophoresis Society July 4–7, Copenhagen, pp. 369–373.
58 Celis, J.E., Madsen, P., Celis, A., Nielsen, H.V. and Gesser, B. (1987) FEBS Lett. 220, 1–7.
59 Moss, B. (1991) Science 252, 1662–1667.
60 Leffers, H., Madsen, P., Honoré, B., Rasmussen, H.H., Andersen, A.H., Walbum, E., Vandekerckhove, J. and Celis, J.E. (1992) (unpublished data).

61 Madsen, P., Rasmussen, H.H., Leffers, H., Honoré, B., Dejgaard, K., Olsen, E., Kiil, J., Walbum, E., Andersen, A., Basse, B., Lauridsen, J., Ratz, G., Celis, A., Vandekerckhove, J. and Celis, J.E. (1991) J. Invest. Dermatol. 97, 701–712.