Minireview

Human 2-D PAGE databases for proteome analysis in health and disease: http://biobase.dk/cgi-bin/celis

Julio E. Celis[a,*], Pavel Gromov[a], Morten Østergaard[a], Peder Madsen[a], Bent Honoré[a], Kurt Dejgaard[a], Eydfinnur Olsen[a], Henrik Vorum[a], Dan Bach Kristensen[a], Irina Gromova[a], Anders Haunsø[a], Josef Van Damme[b], Magda Puype[b], Joel Vandekerckhove[b], Hanne Holm Rasmussen[a]

[a]*Department of Medical Biochemistry and Danish Centre for Human Genome Research, University of Aarhus, Ole Worms Allé, Building 170, DK-8000 Aarhus C, Denmark*
[b]*Lab voor Fysiologische Scheikunde, Ledeganckstraat 35, B-9000 Gent, Belgium*

Received 15 October 1996

Abstract Human 2-D PAGE Databases established at the Danish Centre for Human Genome Research are now available on the World Wide Web (http://biobase.dk/cgi-bin/celis). The databanks, which offer a comprehensive approach to the analysis of the human proteome both in health and disease, contain data on known and unknown proteins recorded in various IEF and NEPHGE 2-D PAGE reference maps (non-cultured keratinocytes, non-cultured transitional cell carcinomas, MRC-5 fibroblasts and urine). One can display names and information on specific protein spots by clicking on the image of the gel representing the 2-D gel map in which one is interested. In addition, the database can be searched by protein name, keywords or organelle or cellular component. The entry files contain links to other databases such as Medline, Swiss-Prot, PIR, PDB, CySPID, OMIM, Methabolic pathways, etc. The on-line information is updated regularly.

Key words: Human genome; Protein identification; Database; Disease

1. Introduction

It is now 6 years since the Human Genome Project was launched and, indeed, there have been remarkable advances both on the construction of physical and genetic maps as well as in identification of genes associated with diseases [1]. Research at the protein level, however, has lagged behind partly due to the complexity of the technology required to separate, analyse and identify the thousands of polypeptides that constitute the proteome[1] of a given human cell type. To date, only a small proportion of the total proteome has been identified and little is known about the protein composition of individual cells, in particular differentiated ones. Proteins orchestrate most cellular functions and therefore, it is of importance to develop technology to resolve and identify them, as well as to pinpoint those associated with diseases.

At present, the only available approach for a comprehensive analysis of the human proteome is provided by high-resolution two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) [2–5] in combination with methods to identify and quantitate the polypeptides present in the individual spots ([6–8] and references therein). The current 2-D PAGE technology allows the resolution of about 3000–4000 [^{35}S]methionine-labelled polypeptides (ranging in pI from 4 to $\approx$12) from any cell line or tissue and of these at least 1000 may correspond to modified variants [6,9,10]. This may seem like a small number compared with the estimated 100 000 expressed human genes [1]; however, only a fraction of the genes are switched on in a given cell type with may be no more than 5000–6000 different expressed proteins (perhaps with the exception of cell types in the brain) and their modifications per cell [9,10]. As much as about 80% of the polypeptides could represent house-keeping proteins that are shared by all cell types and that are expressed in variable amounts ([10] and references therein). Assuming that there are at least 250 different cell types in the human body [11], each expressing about 3–400 proteins unique to a specific cell type, one ends up with a total number of proteins that is reasonable close to the estimated number of genes. Thus, by focusing on the study of common proteins and by co-ordinating the analysis of individual human cell types, it may be possible to resolve and analyse the majority of the expressed proteins in the near future. The purpose of this short review is to describe some features of our 2-D PAGE databases for the study of global cell regulation and diseases (http://biobase.dk/cgi-bin/celis), as well as to underline future directions and limitations of the approach. The reader, however, is encouraged to visit other related sites in the Web, in particular, the federated 2-D PAGE databases that are available individually or through the ExPASy molecular biology server in Geneva (http:/expasy.hcuge.ch/www/expassy-top.html).

2. 2-D gel technology

For the past 20 years, high-resolution 2-D PAGE has been the technique of choice to analyse the protein composition of a given cell type and for monitoring changes in gene activity through the quantitative and qualitative analysis of the thou-

*Corresponding author. Fax: (45) 8613-1160:
E-mail: jec@biokemi.aau.dk

Abbreviations: 2-D, two-dimensional; IEF, isoelectric focusing; NEPHGE, non-equilibrium pH gradient electrophoresis; keywords, sample spot; PAGE, polyacrylamide gel electrophoresis

[1]The term proteome was recently coined by M. Wilkins and K. Williams.

PII S0014-5793(96)01247-1

Fig. 1. Part of a synthetic image of human keratinocyte proteins separated by IEF 2-D PAGE. Proteins flagged with a red cross correspond to known proteins. By clicking on any spot is possible to obtain a file containing information about the protein.

→

sands of proteins that orchestrate various cellular functions. The technique, which was originally described by O'Farrell [2,4] and Klose [3], 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 (relative abundance, post-translational modifications, coregulated proteins) both qualitative and quantitatively in relation to each other.

A systematic analysis of the human proteome by 2-D PAGE requires a reproducible gel system to resolve the proteins as well as computer-assisted technology to scan the gels, make synthetic images, assign numbers to individual spots and match gel spots. In addition, one needs functions to enter and retrieve qualitative and quantitative information. To date, our work has been based entirely on the use of carrier ampholytes as originally described by O'Farrell [2,4]. Even though gels run with carrier ampholytes are difficult to reproduce in many hands we have standardized the technology to a level in which it is possible to obtain reproducible separations routinely ([12], see also procedures and videos at http://biobase.dk/cgi-bin/celis). Recently, Bjellquist and others [13–15] introduced the use of immobilized pH gradients (IPG), which are integral part of the polyacrylamide matrix and provide more reproducible focussing patterns. IPGs avoid some of the problems associated with carrier ampholytes (e.g., cathionic drift) and offer enhanced resolution as very narrow pH gradients of about 0.05 pH/cm can be established.

3. Making a 2-D page database

The first step in making a comprehensive 2-D PAGE database is to prepare a synthetic image (digital form of the gel image) of the gel (autoradiogram, 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 multiwhire chambers [16]. Various softwares for 2-D PAGE analysis have been developed that work on different hardwares ([16] and references therein).

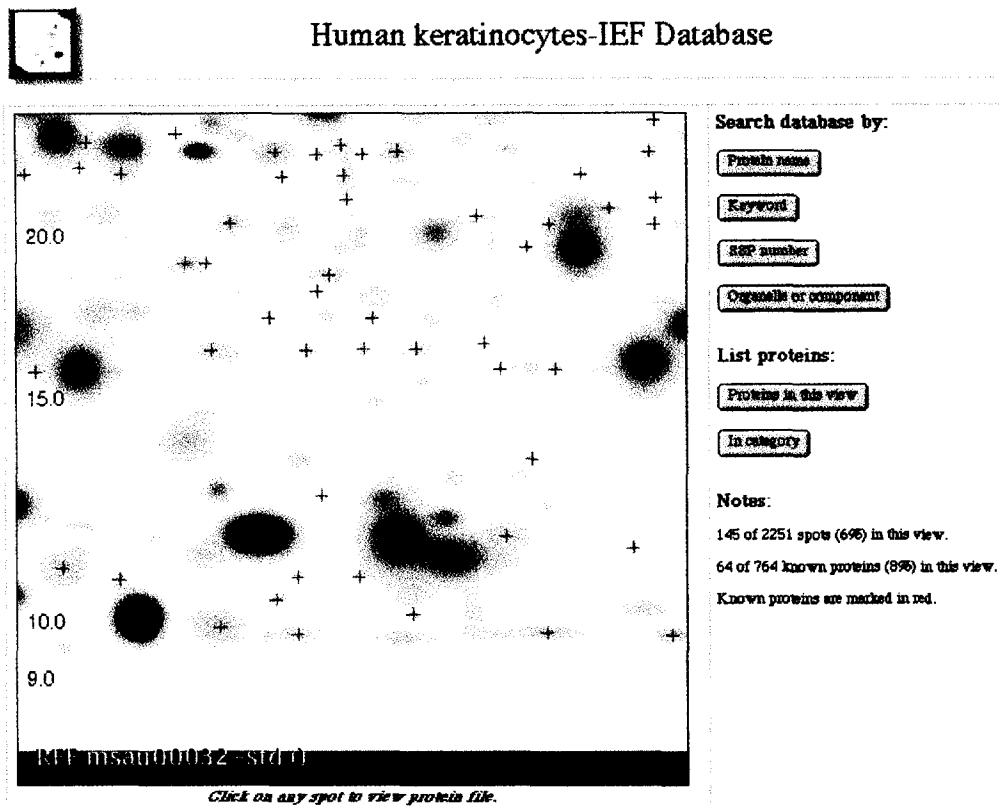
In our workstation in Aarhus, autoradiograms are scanned with a laser scanner and the data are analyzed with the use of the PDQUEST II software (Protein Databases Inc.) [17] running on a SPARK station computer from SUN Microsystems, Inc. The scanner measures intensity in the range of 0 to 3.0 absorbance. A typical scan of a 17×17 cm autoradiogram takes about 2 min. Steps in image analysis include: image acquisition, initial smoothing, background subtraction, final smoothing, spot detection and fitting of ideal Gaussian 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 and takes about 1–1.5 day per image.

Each polypeptide is assigned a number by the computer, a fact that facilitates the entry and retrieval of qualitative and quantitative information for any given spot in the gel [6,10]. The standard image can be matched automatically by the computer to other standard or reference gels provided a few landmarks spots 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 the data. Once a standard map of a given protein sample is made, one can enter qualitative or quantitative information to establish a reference or master database (http://biobase.dk/cgi-bin/celis) ([6,10,18] and references therein). Categories or entries are created so as to gather information on physical, chemical, biochemical, physiological, genetic, architectural as well as biological properties of proteins. In general, entries reflect the type of biological problem that is being studied using the database approach.

4. http://biobase.dk/cgi-bin/celis

As a result of a long-term and systematic effort to study gobal cell regulation and diseases (psoriasis, bladder cancer) we have established several comprehensive, computer-aided IEF and NEPHGE 2-D PAGE databases (non-cultured keratinocytes [10,19], non-cultured transitional cell carcinomas, MRC-5 fibroblasts [20] and urine [21] that can be accessed (in part) on the World Wide Web thanks to a custom made software developed by Protein Databases Inc. (PDI). Of the databanks, the keratinocyte one has taken nearly 6 years to develop and is by far the most comprehensive. The ultimately aim of this database is to gather — through a systematic study of these cells — qualitative and quantitative infomation on keratinocyte proteins that may allow us to identify abnormal patterns of gene expression and to pinpoint signalling pathways and components that are affected in diseases, psoriasis in particular ([21–27] and references therein). Some interesting features of the keratinocyte database, including links to other sites in the Web, are illustrated below.

Fig. 1 shows a fraction of the synthetic image of non-cultured human keratinocyte proteins separated by isoelectric focussing (IEF) 2-D PAGE as depicted on the World Wide Web (http://biobase.dk/cgi-bin/celis). Proteins flagged with a red cross correspond to known polypeptides. To date, about 1100 polypeptides have been identified in this database (IEF and NEPHGE) of the nearly 3200 that have been resolved and catalogued [10]. Proteins have been identified by one or a combination of techniques that include (i) 2-D gel immunoblotting using specific antibodies and the ECL detection procedure [10,28], (ii) microsequencing of Coomassie Brilliant Blue-stained proteins ([29–31] and references therein), (iii) comigration with known human proteins (individual proteins and organelle components), (iv) mass spectrometry of tryptic peptides ([32] and references therein), (v) vaccinia virus [33] and COS-1 expression [34] of full-length cDNAs, (vi) in vitro transcription-translation [34], (vii) blot overlay techniques [35]



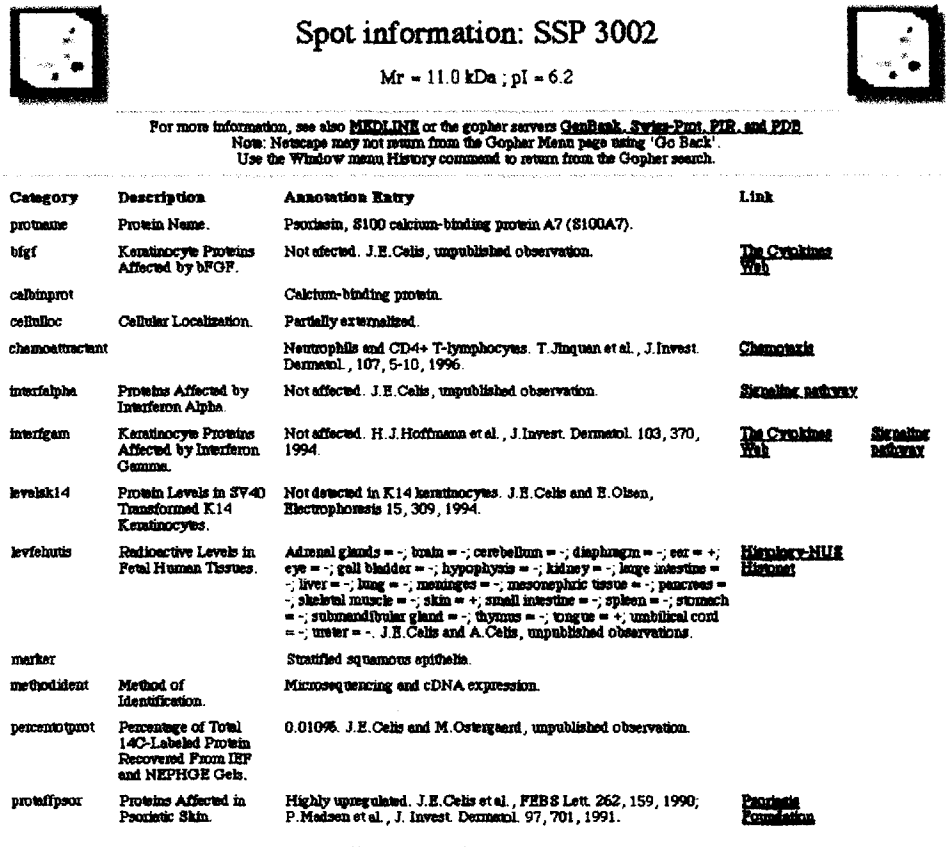


Fig. 2. File for protein SSP 3002 (psoriasin).

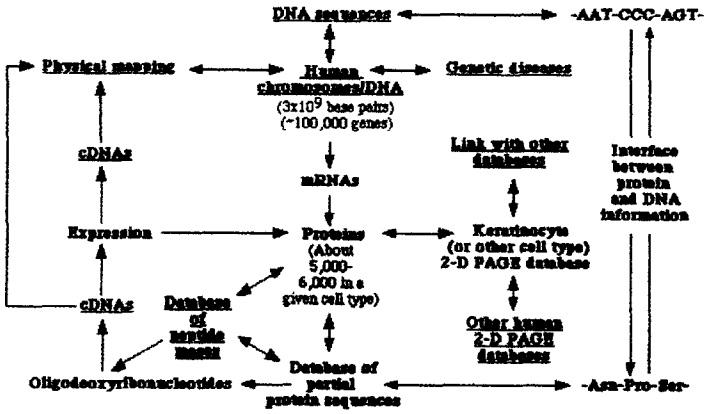


Fig. 3. Linking protein and DNA sequencing and mapping information.

and (viii) labeling of post-translational modifications [36]. Both microsequencing and mass spectrometry are instrumental for the identification of novel proteins.

Information gathered on any given polypeptide, known or unknown, can be easily retrieved by clicking on the corresponding spot. A file containing all of the information gathered for this particular protein (mostly obtained in our laboratories) appears as shown in Fig. 2 (only a fraction of the file for psoriasin is shown). Files for known proteins contain links to a subset of Medline, Swiss-Prot, PIR, and PDB. Other links include OMIM and other Web sites such as CySPID (cytoskeletal protein interactions database), metabolic path-

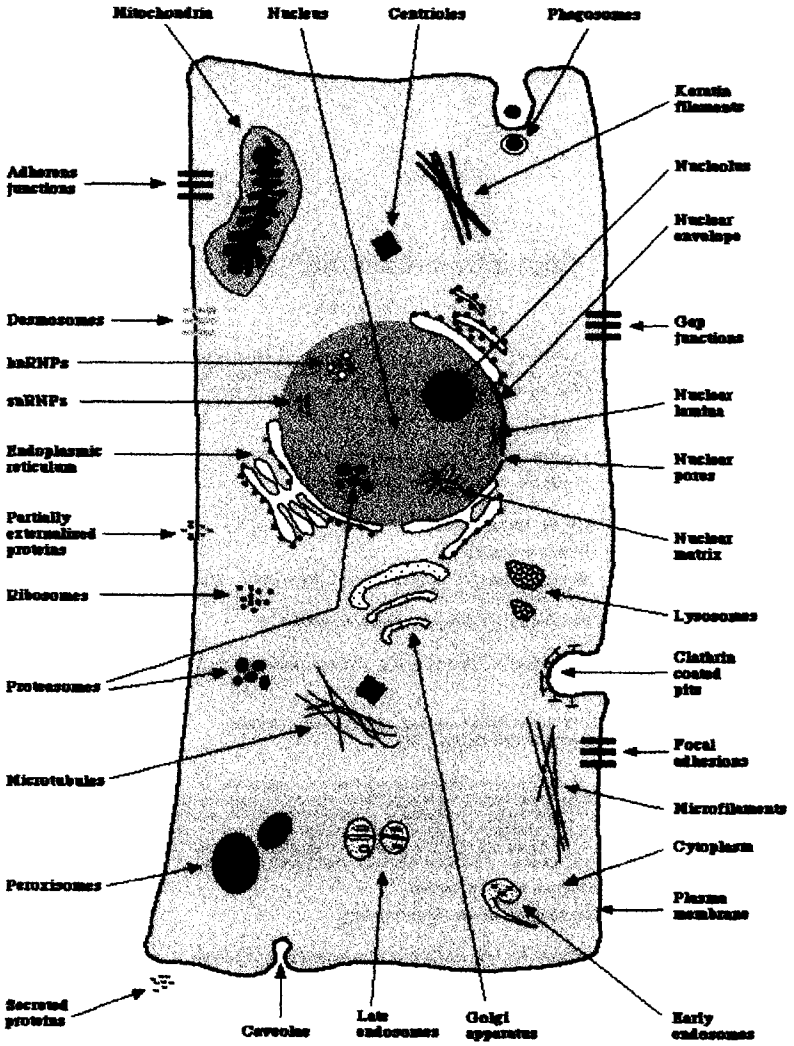


Fig. 4. Organelles and cellular components. By clicking on any of the structures and components it is possible to obtain a protein list as well as their relative positions in the master image.

ways (compiled by KEGG), the cytokines database, NUS histonet (histology), etc. (Fig. 2). In the long run, as new databases and related Web sites become available it will be possible to navigate through various databanks containing complementary information (nucleic acid and protein sequence, genome mapping, diseases, protein structure, post-translational modifications, antibodies, signalling pathways etc.) (Fig. 3). In particular, as the Human Genome Project progress, 2-D PAGE databases are expected to help annotating DNA sequences and will be instrumental in linking protein and DNA sequencing and mapping information. In addition, they may offer a global approach to the study of cell regulation (Fig. 3) [10].

Functions to query the database include search by name or keywords (Fig. 1) as well as by organelle or cellular component (Fig. 4). By clicking on any of the organelles, cellular structures and components depicted in Fig. 3 it is possible to get a protein list (together with M_r and pI) as well as their relative position in the master image. Also, one can retrieve a list of the nearly 1100 known proteins recorded in the database and from there go directly to the position in the gel and thereafter to the files for these proteins. To date, 56 information categories are available on the World Wide Web version of the IEF keratinocyte database. To name a few, these include cellular localization, pathways, proteins affected in psoriatic keratinocytes, proteins differentially regulated in abnormally differentiated keratinocytes, effects of growth factors and cytokines, protein levels in normal and transformed keratinocytes, levels in fetal human tissues, abundancy, distribution in Triton cytoskeletons and supernatans, peptide sequences, cytoskeletal proteins, calcium binding proteins, annexins, chaperonins, heat shock proteins, etc.

The transitional carcinomas databank is being developed with the aim of identifying a complete set of protein biomarkers that may be useful to classified histopathological grades of TCCs and that will provide specific probes (mainly antibodies) for the objective diagnosis, prognosis and treatment of these lesions [37,38]. In due course, these biomarkers will be used as landmarks for forthcoming research directed at dissecting the various stages involved in tumour progression. Furthermore, we are interested in identifying tumour specific antigens that are externalized to the urine [21,37], and that may identify individuals at risk.

4.1. Concluding remarks and perspectives

Recent advances in protein identification techniques (microsequencing, mass spectrometry immunoblotting and cDNA expression systems) have made possible the establishment of comprehensive 2-D PAGE gel protein databases that aim at linking protein and DNA mapping and sequence information (Human Genome Project), and that offer a unique approach to the study of the proteome both in health and disease (see also http:/expasy.hcuge.ch/www/expassy-top.html). With the integrated approach offered by 2-D PAGE databases it is now possible to reveal phenotype-specific proteins, sequence them, to search for homology with previous identified proteins, clone the cDNAs, 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 co-ordinately expressed in a given biological process, diseases included. Human 2-D PAGE 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 complement the current efforts to map and sequence the entire human genome. In addition, 2-D PAGE databases and associated technology are expected to address problems that cannot be approached by DNA analysis, namely, relative abundancy of the protein product, post-translational modifications, subcellular localization, turnover, interaction with other proteins as well as functional aspects.

In spite of new and sensitive technology available for proteome analysis, in particular the development of the nano-electrospray by Wilm and Mann [39] that allows the identification of proteins in the low picomole level, we believe that a complete analysis of the human proteome will be hampered by the lack of procedures to prepare pure non-cultured normal cells from various tissues. This may seem trivial at first glance, but many normal cells may change their protein expression profile shortly upon culturing given a distorted view of the true expression pattern. Working with whole tissues on the other hand, poses many technical problems due to cell heterogeneity and high abundancy of some differentiated products. Obviously, one can enrich for low abundance proteins by cell fractionation or conventional protein purification procedures, but the more one handles a sample the higher the chance of introducing artificial protein modifications. We also foresee problems with the separation of very basic proteins, the identification of very low abundance proteins that are post-translationally modified as well as with quantitations in general as none of the current procedures can be used to analysize all of the proteins resolved in a gel. All in all however, the 2-D PAGE technology and satellite techniques have now reached a stage in which it is possible to begin to contemplate translating basic discoveries into meaningful clinical applications.

Acknowledgements: We would like to thank I. Andersen, B. Basse, A. Celis, J.B. Lauridsen and G. Ratz for expert technical assistance. This work was supported by grants from the Danish Biotechnology Program, The Danish Cancer Society, NOVO, The Aarhus University Research Fund, The European Union and the Medical Society for the Advancement of Science.

References

[1] The genome directory (1995) Nature 377 (6547 Suppl.), 1–379.
[2] O'Farrell, P.H. (1975) J. Biol. Chem. 250, 4007–4021.
[3] Klose, J. (1975) Humangenetik 26, 231–243.
[4] O'Farrell, P.Z., Goodman, H.M. and O'Farrel, P. (1977) Cell 12, 1133–1141.
[5] Celis, J.E. and Bravo, J., Eds. (1984) Two-Dimensional Gel Electrophoresis of Proteins: Methods and Applications. Academic Press, New York.
[6] 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.
[7] Wilkins, M.R., Sanchez, J.-C., Gooley, A.A., Appel, R.D., Humphery-Smith, I., Hochstrasser, D.F. and Williams, K.L. (1995) Biotech. Gene. Eng. Rev. 13, 19–50.
[8] Lottspeich, F. (Ed.) (1996) Special Issue: Electrophoresis 5, 811–966.
[9] 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., Vandekerckhove, J. (1990) Electrophoresis 11, 989–1071.
[10] Celis, J.E., Rasmussen, H.H., Gromov, P., Olsen, E., Madsen, P.,

Leffers, H., Honoré, B., Dejgaard, K., Vorum, H., Kristensen, D.B., Østergaard, M., Haunsø, A., Jensen, N. Aa., Celis, A., Basse, B., Lauridsen, J.B., Ratz, G.P., Andersen, A.H., Walbum, E., Kjærgaard, I., Andersen, I., Puype, M., Van Damme, J., Vandekerckhove, J. (1995) Electrophoresis 12, 2177–2240.
[11] Genrser, F. (1986) Textbook of Histology. Munksgaard, Copenhagen.
[12] Celis, J.E., Ratz, G., Basse, B., Lauridsen, J.B. and Celis, A. (1994) in: Cellbiology: a Laboratory Handbook (Celis, J.E. Ed.), Vol. 3, pp. 222–230, Academic Press, San Diego, CA.
[13] Bjellqvist, B., Ek, K., Righetti, P.G., Gianazza, E., Görg, A., Westermeier, R. and Postel, W. (1982) J. Biochem. Biophys. Methods 6, 317–339.
[14] Görg, A., Postel, W. and Günther, S. (1988) Electrophoresis 9, 531–546.
[15] Righetti, P.G. (1990) Immobilized pH Gradients: Theory and Methodology. Elsevier, Amsterdam.
[16] Miller, M.J. (1989) Adv. Electrophoresis 3, 182–217.
[17] 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, New York.
[18] Celis, J.E. (Ed.) (1990) Special issue: two-dimensional gel protein databases, Electrophoresis 11.
[19] Celis, J.E., Madsen, P., Rasmussen, H.H., Leffers, H., Honoré, B., Gesser, B., Dejgaard, K., Olsen, E., Magnusson, N., Kiil, J., Celis, A., Laurisen, 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., Vandekerckhove, J. (1991) Electrophoresis 12, 802–872.
[20] Celis, J.E., Dejagaard, 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 Vandkerckhove, J. (1990) Electrophoresis 11, 1072–1113.
[21] Rasmussen, H.H., Ørntoft, T.F., Wolf, H. and Celis, J.E. (1996) J. Urol. 155, 2113–2119.
[22] Madsen, P., Rasmussen, H.H., Leffers, H., Honoré, B., Dejgaard, K., Olsen, E., Kiil, J., Walbum, E., Andersen, A.H., Basse, B., Lauridsen, J.B., Ratz, G.P., Celis, A., Vandekerckhove, J. and Celis, J.E. (1991) J. Invest. Dermatol. 97, 701–712.
[23] Madsen, P., Rasmussen, H.H., Leffers, H., Honoré, B. and Celis, J.E. (1992) J. Invest. Dermatol. 99, 299–305.
[24] Rasmussen, H.H. and Celis, J.E. (1993) J. Invest. Dermatol. 101, 560–566.
[25] Madsen, P., Rasmussen, H.H., Flint, T., Gromov, P., Kruse, T.A., Honoré, B., Vorum, H. and Celis, J.E. (1995) J. Biol. Chem. 270, 5823–5829.
[26] Olsen, E., Rasmussen, H.H. and Celis, J.E. (1995) Electrophoresis 19, 2241–2248.
[27] Jinquan, T., Vorum, H., Grønhøj Larsen, C., Madsen, P., Rasmussen, H.H., Gesser, B., Etzerodt, M., Honoré, B., Celis, J.E and Thestrup-Pedersen, K. (1996) J. Invest. Dermatol. 107, 5–10.
[28] Celis, J.E., Lauridsen, J.B. and Basse, B. (1994) in: Cell Biology: a Laboratory Handbook (Celis, J.E. ed.), Vol. 2, pp. 305–313, Academic Press, San Diego, CA.
[29] Bauw, G., Van Damme, J., Puype, M., Vandekerckhove, J., Gesser, B., Ratz, G.P., Lauridsen, J.B. and Celis, J.E. (1989) Proc. Natl. Acad. Sci. USA 86, 7701–7705.
[30] Rasmussen, H.H., Van Damme, J., Bauw, G., Puype, M., Celis, J.E. and Vandekerckhove, J. (1991) in: Methods in Protein Sequence Analysis (Jörnvall, H., Höög, J.-O. and Gustavsson, A.-M. eds.), pp. 103–104, Birkhäuser, Basel.
[31] Vandekerckhove, J. and Rasmussen, H.H. (1994) in: Cell Biology: a Laboratory Handbook (Celis, J.E. ed.), Vol. 3, pp. 359–368, Academic Press, San Diego, CA.
[32] Rasmussen, H.H., Mørtz, E., Mann, M., Roepstorff, P. and Celis, J.E. (1994) Electrophoresis 15, 406–416.
[33] Leffers, H., Madsen, P., Rasmussen, H.H., Honoré, B., Andersen, A.H., Walbum, W., Vandekerckhove, J. and Celis, J.E. (1993) J. Mol. Biol. 231, 982–998.
[34] Madsen, P., Gromov, P. and Celis, J.E. (1995) Electrophoresis 16, 2258–2261.
[35] Gromov, P.S. and Celis, J.E. (1994) Electrophoresis 15, 474–481.
[36] Gromov, P.S. and Celis, J.E. (1997) in: Cell Biology: a Laboratory Handbook (J.E. Celis, ed.), Academic Press, New York, in press.
[37] Celis, J.E., Rasmussen, H.H., Vorum, H., Madsen, P., Honoré, B., Wolf, H. and Ørntoft, T.F. (1996) J. Urol. 155, 2105–2112.
[38] Celis, J.E., Østergaard, M., Basse, B., Celis, A., Lauridsen, J.B., Ratz, G.P., Andersen, I., Hein, B., Wolf, H., Ørntoft, T.F. and Rasmussen, H.H. (1996) Cancer Res. 15, 4782–4790.
[39] Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T. and Mann, M. (1996) Nature 79, 466–469.