Perspectives

Comprehensive, human cellular protein databases and their implication for the study of genome organization and function

Julio E. Celis, Gitte P. Ratz, Peder Madsen, Borbala Gesser, Jette B. Lauridsen, Sianette Kwee, Hanne Holm Rasmussen, Henrik V. Nielsen, Dorthe Crüger, Bodil Basse, Henrik Leffers, Bent Honoré, Olaf Møller, Ariana Celis, Joel Vandekerckhove+, Guy Bauw+, Jozef van Damme+, Magda Puype+ and Marc Van den Bulcke+

Institute of Medical Biochemistry and Bioregulation Research Centre, Aarhus University, DK-8000 Aarhus C, Denmark and +*Laboratorium voor Genetica, Rijksuniversiteit Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium*

Received 27 December 1988

Comprehensive, computerized databases of cellular protein information derived from the analysis of two-dimensional gels, together with recently developed techniques to microsequence proteins offer a new dimension to the study of genome organization and function. In particular, human protein databases provide an ideal framework in which to focus the human genome sequencing effort.

Protein database; Microsequencing; Genome organization; Genome function; Cloning; Genome sequencing

1. INTRODUCTION

Since its development in 1975 [1], high resolution 2D gel electrophoresis has provided a unique tool to examine the protein composition and overall patterns of gene expression of a given cell type ([2,3] and references therein). About 2000 proteins can be separated using this technology, a number that corresponds to about 2 million base pairs of coded DNA. Much of the information generated so far however, has not reached the scientific community in its fullness, and only recently thanks to the development of appropriate computer software (reviewed in [4]), it has been possible to scan, assign numbers to individual polypeptides, compare, quantitate and store the wealth of information contained in the gels. This important development has allowed us to make full use of existing data, and to construct comprehensive, computerized databases of cellular protein information ([3–8] and references therein). Human protein databases [6,7] are becoming increasingly important in view of the concerted effort to map and sequence the entire human genome. Some features of these databases as well as recent technical developments concerning protein microsequencing are briefly discussed below.

2. COMPRENSIVE, COMPUTERIZED DATABASES OF HUMAN CELLULAR PROTEINS

To date, three databases of human cellular proteins (transformed epithelial amnion cells (AMA)

Correspondence address: J.E. Celis, Institute of Medical Biochemistry and Bioregulation Research Centre, Aarhus University, DK-8000 Aarhus C, Denmark

Abbreviations: 2D, two-dimensional; IEF, isoelectric focusing; NEPHGE, non-equilibrium pH gradient electrophoresis; PCNA, proliferating cell nuclear antigen; PTH, phenylthiohydantoin; TFA, trifluoroacetic acid

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/89/$3.50 © 1989 Federation of European Biochemical Societies

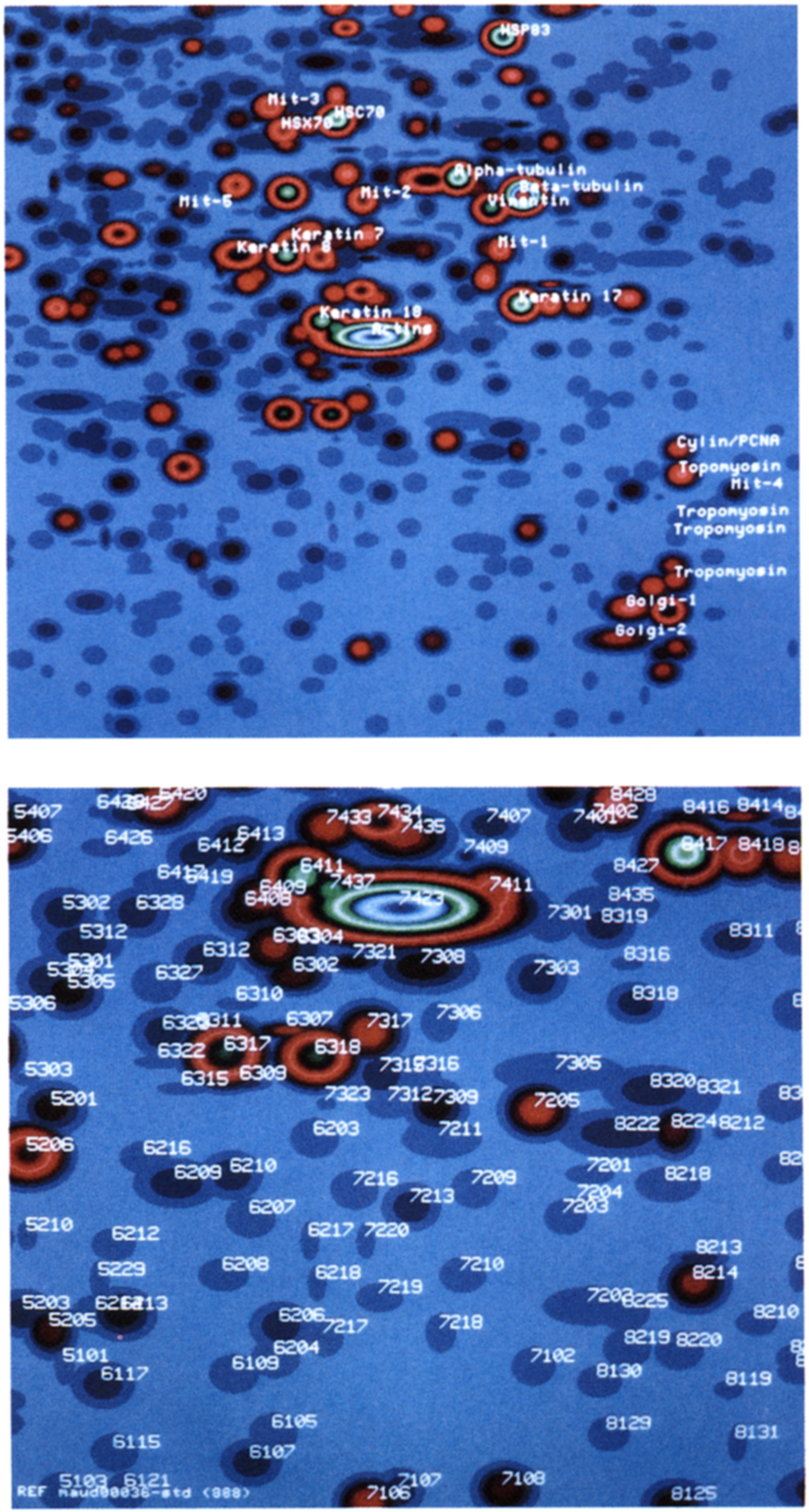

Fig.1. Synthetic images of IEF gels of [^{35}S]methionine-labelled AMA cellular proteins showing protein names (upper panel) and numbers (lower panel). Only a fraction of the gel is shown.

[6], peripheral blood mononuclear cells (PBMC) [6], and embryonic lung MRC-5 fibroblasts [7]) have been published in a comprehensive form. Some important features of these databases are exemplified here using data from the master AMA database [6]. In this database a total of one thousand seven hundred and eighty-two [^{35}S]methionine-labelled polypeptides have been separated (fig.1, upper panel; only a fraction of the IEF synthetic image is shown) and recorded using computer analyzed two-dimensional gel electrophoresis (PDQUEST) [9]. Each polypeptide has been assigned a number (fig.1, lower panel), and various categories have been created to enter qualitative (annotations, fig.1, upper panel) and quantitative data (fig.2, quantitations of tropomyosins in normal and SV40 transformed human MRC-5 fibroblasts) [7]. In general, categories or entries are created so as to compile information concerning physical, chemical, biochemical, physiological, genetic, architectural and biological properties of proteins. Table 1 gives an example of some of the entries available for the cell cycle regulated and DNA replication protein cyclin/PCNA (see also fig.1, upper panel) [6]. This proliferation-sensitive nuclear protein was one of the first interesting polypeptides identified using the 2D gel technology [10].

Clearly, computerized, comprehensive databases allow easy access to a large body of data and provide an efficient medium to communicate standardized protein information. Once a protein is identified in a database, all of the information accumulated, can be easily retrieved and made available to the researcher.

Databases will evolve as more proteins are identified (comigration with purified proteins, immunoblotting using specific antibodies, compari-

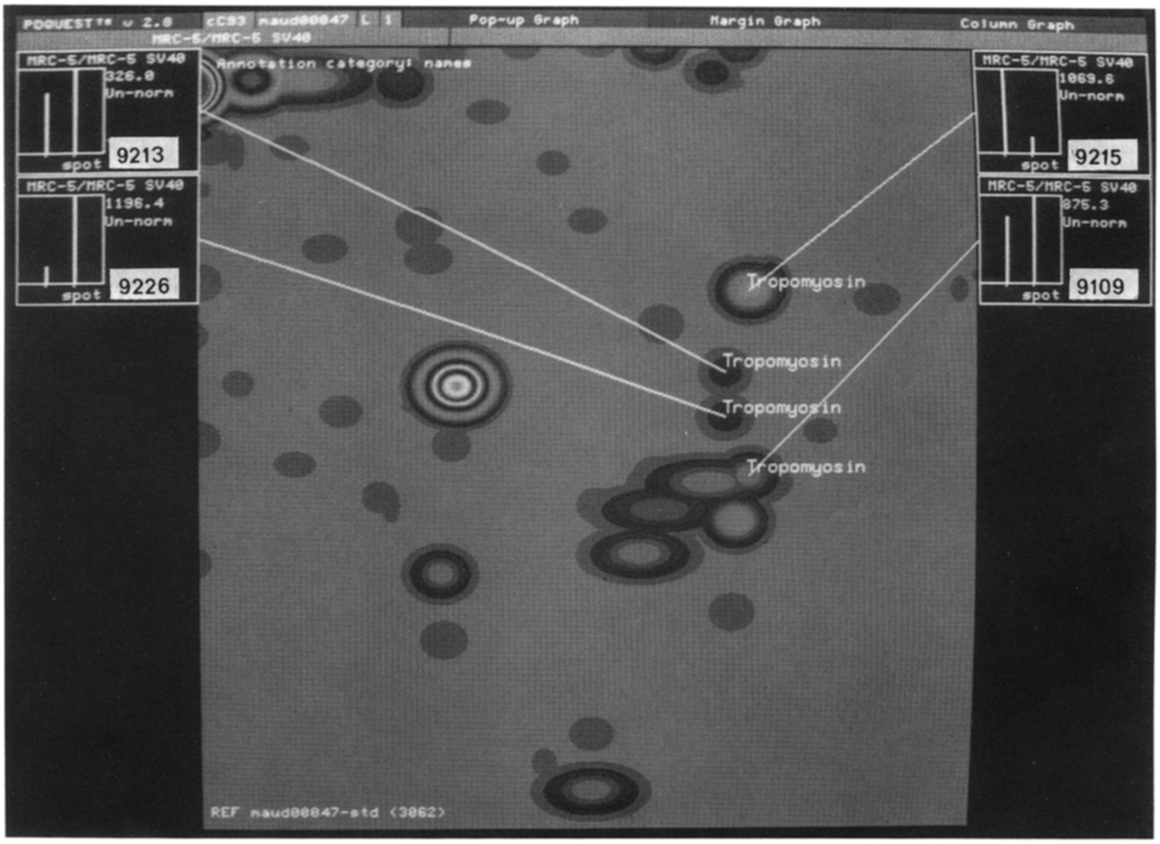

Fig.2. Synthetic image of a fraction of an IEF fluorogram of [^{35}S]methionine-labelled proteins from normal human embryonic lung MRC-5 fibroblasts (p35). The histograms show levels of synthesis of tropomyosins in MRC-5 (left bar) and MRC-5 SV40 (MRC5-V2) (right bar) human embryonal lung fibroblasts. The height of the bars is scaled to the maximum detected value.

Table 1

Entries for cyclin/PCNA in the human AMA protein database[a]

Entries for cyclin/PCNA Protein 9218	Information entered
DNA sequence	Known [24]
Amino acid sequence	Known. Deduced from DNA sequence [24]
Subcellular localization	Nuclear. Sites of DNA replication [25]. Nuclear patterns of cyclin/PCNA antigen staining subdivide S phase [26]
Function	Some step in DNA replication [25,27]. Coordinated expression of leading and lagging strand [28]. Also a role in DNA repair [29]
Phosphorylation	Not phosphorylated [6]
Synthesis in proliferating and non-proliferating cells	Synthesized mainly by proliferating cells of both normal and transformed origin ([30] and references therein)
Cell cycle specificity	Preferentially synthesized during S-phase. Increased synthesis starts at the G_1/S transition border [10]
Presence in various vertebrate species	Present in aves, bat, dog, dolphin, goat, hamster, human, mink, monkey, mouse, pisces, potoroo, rabbit and rat ([31] and references therein)
Coregulated proteins	A few have been identified by Garrels and Franza in rat REF52 cells [8]
Variants	Two. One more acidic variant present in hamster, human, potoroo and rat, and the other more basic present in mouse ([31] and references therein)
Presence in adult human tissues	Not detected in the following tissues: aorta, bladder, cavum oris, cerebellar cortex, cerebral cortex, cornea, ductus deferens, epididymis, fat tissue, heart muscle, kidney cortex, kidney medulla, larynx, lung, mammary gland, medulla oblongata, mesencephalon, palate, pharynx, posterior eye polus, prostata, rectum, skeletal muscle, skin, submaxillary glands, thyroid gland, trachea, ureter, uterus, veins, and vesicular seminalis [32]
Presence in fetal human tissues	Small intestine, thymus, spleen, kidney, liver and lung exhibit high levels of cyclin/PCNA ([6], see also [33] for data on mouse tissues)
Antibodies produced against protein	Monoclonals [34] and rabbit polyclonal antibodies ([35], unpublished observations of Gesser, B., Basse, B. and Celis, J.E.)
Antibodies detected in human sera	In a few percentages of patients with Systemic Lupus Erythematosus [36]
Associated proteins	Associated with DNA polymerase δ [35,37,38]
Cell type in which highest expression has been observed so far	Human Molt-4 cells [30]
Purification procedure	Available in article by Ogata et al. [39]
Effects of drugs	Cyclin/PCNA synthesis is not affected by aphidicolin or hydroxyurea [40]
HeLa protein catalogue number	IEF 49 [41]
Mouse protein catalogue number	IEF 51f [42]

[a] Modified from Celis et al. [6]

son of peptide sequences to sequences stored in protein databanks (see below)), and new information, gathered worldwide, is entered in the form of annotations. Colleagues are encouraged to send us samples of purified human proteins and/or antibodies (broad species specificity) that could be used to identify new proteins.

Comprehensive human databases may become important tools for the study of diseases, and may provide an ideal framework in which to focus the human genome mapping and sequencing effort.

3. MICROSEQUENCING HAS ADDED A NEW DIMENSION TO COMPREHENSIVE HUMAN PROTEIN DATABASES

One of the major problems encountered in building databases based on the analysis of two-dimensional gels is the identification and characterization of the large number of proteins separated by this technology. Here, protein microsequencing may come into play. Indeed, generated pieces of protein sequences can be compared with available sequences stored in databanks in order to identify known proteins.

The development of highly sensitive amino acid gas-phase or liquid-phase sequenators [11], together with the establishment of very efficient protein and peptide sample preparation methods, have opened the possibility to carry out a systematic sequence analysis of proteins resolved by 2D gel electrophoresis. Already in the early seventies, gel electrophoresis was used to purify proteins for sequencing purposes (reviewed by Weber and Osborn in [12]). Here, proteins were recovered by diffusion and sequenced by the manual dansyl-Edman degradation (already!) at the nanomole (± 50 μg) level. This new technique was further refined by using electro-elution to recover proteins and by miniaturizing the system [13]. 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 became gradually smaller (e.g. at the 10 picomole level).

Most of these problems have been minimized with the introduction of protein-electroblotting procedures [14–18]. When proteins are blotted on chemically inert membranes, it is possible to sequence the immobilized proteins directly without additional manipulations. Thus, depending on the amount of bound protein and its nature, this direct sequencing procedure generally yields NH_2-terminal sequences containing 10–40 residues. As such, this technique was employed to identify, by their NH_2-terminal sequences, differentially expressed major proteins from total cellular extracts separated on 2D gels [15,16]. Although the method was technically simple and therefore very popular, many laboratories have faced unacceptably high levels of artefactual NH_2-terminal blocking with gel separated proteins. By carrying out gel electrophoresis with buffers of lower pH, initial sequencing yields could be improved, suggesting that the NH_2-terminal blockage takes place during gel electrophoresis, most likely, due to reaction with unpolymerized acrylamide [19]. In addition to these technical problems, many proteins are blocked in vivo by acylation or by a pyrrolidone carboxylic acid cap.

The problem of NH_2-terminal blocking can be circumvented by cleaving the protein and generating internal sequences. This can be accomplished by fragmenting the protein while present in the gel (gel in situ cleavage), or by cleaving it while bound to the membrane (membrane in situ cleavage) [20–22]. In both cases, proteins are either cleaved in a restricted way (e.g. by limited enzymatic digestion or by using restriction chemical cleavage conditions), or fragmented into smaller peptides. In the first case, the large fragments obtained are separated again by gel-electrophoresis, but in the second situation, generated peptides are too small for gel electrophoresis and must be purified by reverse-phase HPLC.

Of the different combinations examined, we experienced very good results by using exhaustive proteolytic digestion on membrane-immobilized proteins. This method has been described for Ponceau red-stained proteins on nitrocellulose blots [21], for Amido-black-stained Immobilon-bound proteins, and for fluorescamine-detected proteins on glass-fiber membranes [22]. The proteases used (trypsin, chymotrypsin 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 HPLC before being sequenced individually. Although each of these manipulations could be expected to result in a reduced yield of final sequence information, we noticed much to our surprise, that the peptides could be sequenced with high efficiency. In our hands, this approach could be routinely applied to gel purified proteins available in amounts ranging from 5 to 10 μg, and often yielded sequence information covering more than 30%

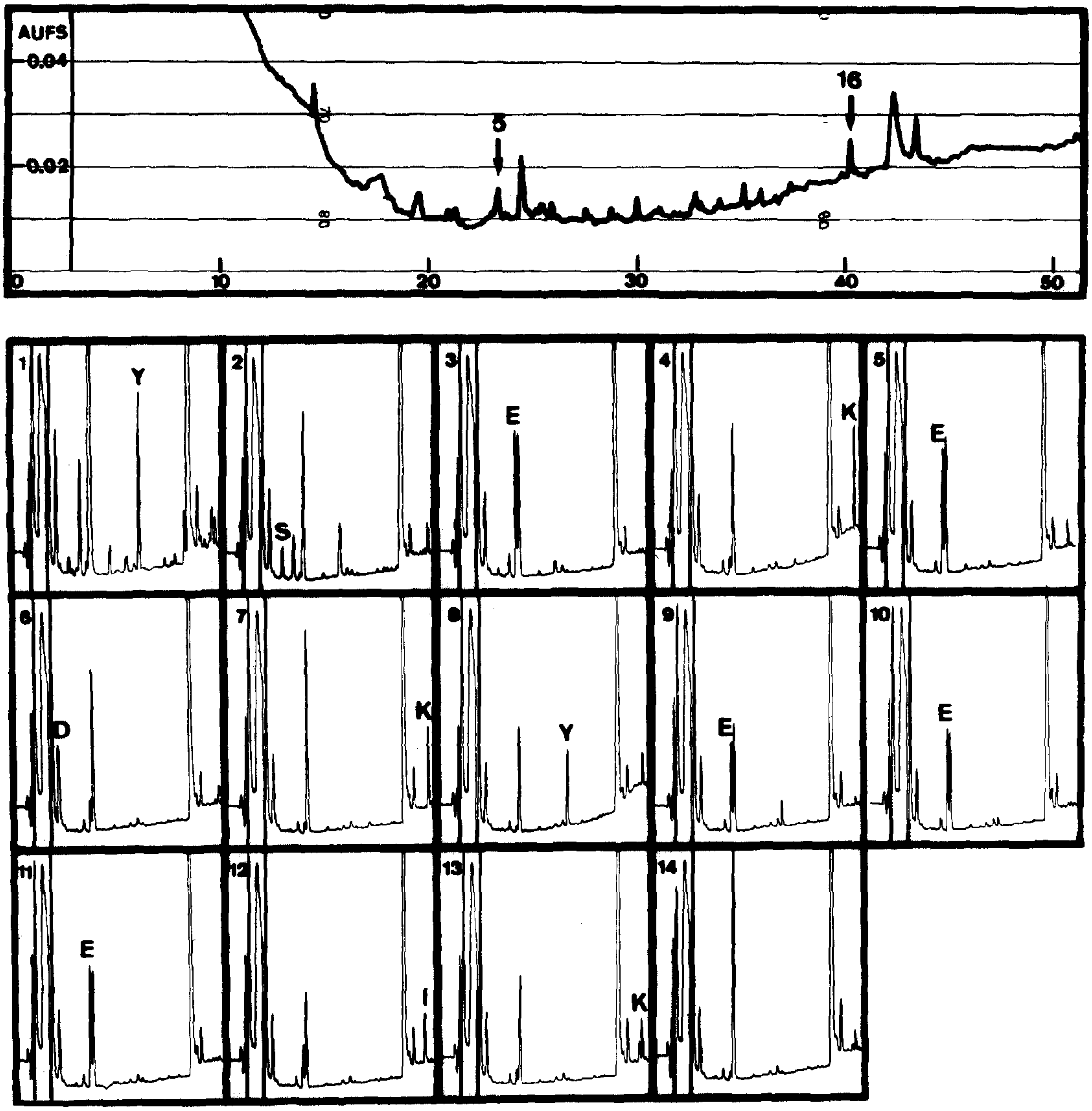

Fig.3. Internal sequences identifying protein IEF9109 [6,7] as a human non-muscle tropomyosin. Tryptic peptides obtained by membrane in situ cleavage were separated on a C4 (4.6 × 250 mm) reverse-phase column (Vydac Separations Group, USA). Solvent A consisted of 0.1% TFA, and solvent B was 70% acetonitrile in 0.1% TFA. A gradient from 0% solvent B to 100% solvent B was applied in a linear mode over 70 min. The eluate (1 ml/min) was recorded by absorbancy at 214 nm (0.2 AUFS) (upper panel). 0 indicates the start of the chromatogram. Peptides were collected manually in Eppendorf tubes. Those indicated by arrows were subjected to gas-phase sequence analysis using a 470 A Applied Biosystems Inc. (USA) gas-phase sequenator equipped with a 120 A on-line PTH-amino acid analyzer. As an illustration, the PTH-residue chromatograms of cycles 1–15 of peak number 5 are shown. Assigned residues are indicated by the one-letter notation (lower panel).

of the total protein. 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 theoretically expected from the number of potential cleavage sites. Consequently, HPLC peptide chromatograms are less complex and the majority of the peptides can be recovered in pure form.

As only limited amounts of a protein mixture can be loaded on a 2D gel, proteins of interest are often obtained in yields which are 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 prior to gel analysis.

Alternatively, spots cut from several gels can be combined and the protein re-eluted. In the example shown in fig.3, protein IEF9109 (see also fig.2) [6,7] was eluted from 10 gel pieces cut from 2D gels of human Molt-4 proteins. The gels had been stained with Coomassie blue according to standard procedures and dried. In this form, gel pieces can be stored for several months or even mailed to laboratories having amino acid sequence facilities. The gel pieces were then further processed by allowing them to re-swell in buffer and placed in a slot (1.5 mm thick, 6 mm broad and 30 mm deep) of a 1D gel. Using the stacking potential of the discontinuous gel system, the protein was eluted and re-concentrated as a sharp band in the second gel. After electroblotting onto membranes, the protein was digested with trypsin and the peptides were separated by reverse-phase HPLC (fig.3, upper panel) and collected individually for sequence determination on a gas-phase sequenator. Comparison of sequence data from two of the peptides (the sequences of peptides 5 and 16 were YSEKEDKYEEEIK [fig.3, lower panel] and EENVGLHQTLDQTLNELNXI, respectively), with sequences stored in the PIR databank, indicated that this protein corresponds to a tropomyosin (aligning with residues 177–189 and 228–247 of human fibroblast tropomyosin). This is in agreement with previous studies which showed that this protein is immunoprecipitated by a monoclonal antibody that reacts with several human tropomyosins [23].

In summary, the membrane in situ proteolytic degradation technique is particularly adapted for sequence analysis of picomolar amounts of proteins and exhibits several advantages over other techniques able to generate internal sequences: (i) proteins are digested for a short period of time with a small amount of protease, thus limiting the formation of protease-derived peptides due to autodigestion and (ii) peptides that originate from protease-sensitive regions in the protein are obtained in high yields and can be sequenced with high efficiency. As far as we can judge, this method avoids most of the drawbacks associated with alternative procedures, and may be considered as the most suitable for obtaining pieces of internal sequences for protein comparison, or for isolation and sequencing of the corresponding cDNA clones. All available information indicates that it may be possible to obtain partial sequence information from most of the proteins resolved by 2D gels that can be visualized by Coomassie blue staining. Indeed, several proteins (catalogued in the AMA protein database) recovered from 2D gels have been microsequenced so far using this procedure (Bauw et al., in preparation).

4. FROM PROTEINS TO CLONED GENES: HUMAN DATABASES AND THEIR IMPLICATION FOR THE STUDY OF GENE ORGANIZATION AND FUNCTION

Comprehensive human protein databases offer an inventory of thousands of proteins many of which are of unknown function. As more proteins are identified, and new information is gathered worldwide, databases will display overall patterns of gene expression and will encourage a global approach to the cell. Most importantly, databases may pinpoint individual or groups of polypeptides (coregulated proteins) that may exhibit interesting regulatory patterns and that may help to focus the human genome mapping and sequencing effort.

In our laboratories, work will be mainly concentrated on the study of proliferation-sensitive proteins ([7] and references therein) as well as cytoskeletal components associated with actin. In the first instance, oligonucleotide probes prepared from information obtained from partial peptide sequences, as well as specific antibodies will be used to clone the cDNAs. These in turn could be used to quantitate message expression, and to screen genomic libraries in an effort to identify regulatory sequences that may be present in groups of proteins that are coregulated.

In the long run, as human genome sequence data will progressively become available, it will also be possible to assign partial protein sequences to genes for which the full DNA sequence and the chromosomal location is known. In this context, microsequencing of 2D gel purified proteins may function as the indispensable link between the 2D gel cellular protein databases and information derived from systematic sequencing of the human genome.

Acknowledgements: We would like to thank S. Himmelstrup Jørgensen for typing the manuscript and O. Sønderskov for photography. The research in Aarhus was supported by grants from the Biotechnology program, the Danish Cancer Foundation, the Danish Medical Research Council, FTU, the Danish Rheumatoid Society, NOVO, the Aarhus University Research Fund and the Fund of Lægevidenskabens Fremme. The work in Gent was supported by grants from the Belgian National Fund for Scientific Research (NFWO) and the Commission of European Communities. J.V. is a Research Associate of the NFWO and G.B. and M.V.d.B. are bursars of the Belgian IWONL.

REFERENCES

[1] O'Farrell, P.H. (1975) J. Biol. Chem. 250, 4007–4021.

[2] Celis, J.E. and Bravo, R. (1984) in: Two Dimensional Gel Electrophoresis of Proteins: Methods and Applications (Celis, J.E. and Bravo, R. eds) Academic Press, New York.

[3] Special Issue of Electrophoresis, February 1989, in press.

[4] Lemkin, P.F. and Lester, E.P. (1989) Electrophoresis, in press.

[5] Phillips, T.D., Vaughn, V., Bloch, P.L. and Neidhardt, F.C. (1987) in: *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology (Neidhardt, F.C. et al. eds) pp.919–966, American Society for Microbiology, Washington, DC.

[6] Celis, J.E., Ratz, G.P., Celis, A., Madsen, P., Gesser, B., Kwee, S., Madsen, P.S., Nielsen, H.V., Yde, H., Lauridsen, J.B. and Basse, B. (1988) Leukemia 9, 561–601.

[7] 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., Møller, O. and Celis, A. (1989) Electrophoresis, in press.

[8] Garrels, J.I. and Franza, B.R., jr (1989) J. Biol. Chem., in press.

[9] Garrels, J.I., Farrar, J.T. and Burwell, I.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.

[10] Bravo, R. and Celis, J.E. (1980) J. Cell Biol. 84, 795–802.

[11] Hewick, R.M., Hunkapiller, M.W., Hood, L.E. and Dreyer, W.J. (1981) J. Biol. Chem. 256, 7990–7997.

[12] Weber, K. and Osborn, M. (1975) in: The Proteins (Neurath, H. et al. eds) pp.179–223, Academic Press, New York.

[13] Hunkapiller, M.W., Lujan, E., Ostrander, F. and Hood, L.E. (1983) Methods Enzymol. 91, 227–236.

[14] Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J. and Van Montague, M. (1985) Eur. J. Biochem. 152, 9–19.

[15] Aebersold, R.H., Teplow, D.B., Hood, L.E. and Kent, S.B.H. (1986) J. Biol. Chem. 261, 4229–4238.

[16] Bauw, G., De Loose, M., Inzé, D., Van Montagu, M. and Vandekerckhove, J. (1987) Proc. Natl. Acad. Sci. USA 84, 4806–4810.

[17] Matsudaira, P. (1987) J. Biol. Chem. 262, 10035–10038.

[18] Eckerskorn, C., Mewes, W., Goretzki, H. and Lottspeich, F. (1988) Eur. J. Biochem. 176, 509–519.

[19] Moos, M., jr, Nguyen, N.Y. and Liu, T.-Y. (1988) J. Biol. Chem. 263, 6005–6008.

[20] Kennedy, T.E., Gawinowicz, M.A., Barzilai, A., Kandel, E.R. and Sweatt, J.D. (1988) Proc. Natl. Acad. Sci. USA 85, 7008–7112.

[21] Aebersold, R.H., Leavitt, J., Saavedra, R.A., Hood, L.E. and Kent, S.B.H. (1987) Proc. Natl. Acad. Sci. USA 84, 6970–6974.

[22] Bauw, G., Van den Bulcke, M., Van Damme, J., Puype, M., Van Montagu, M. and Vandekerckhove, J. (1988) J. Prot. Chem. 7, 194–196.

[23] Celis, J.E., Gesser, B., Small, J.V., Nielsen S. and Celis, A. (1986) Protoplasma 135, 38–49.

[24] Almendral, J.M.D., Huebsch, D., Blundell, P.A., McDonald-Bravo, H. and Bravo, R. (1987) Proc. Natl. Acad. Sci. USA 84, 1575–1579.

[25] Madsen, P. and Celis, J.E. (1985) FEBS Lett. 193, 5–11.

[26] Celis, J.E. and Celis, A. (1985) Proc. Natl. Acad. Sci. USA 82, 3262–3266.

[27] Prelich, G., Kostura, M., Marshak, D.R., Mathews, M.B. and Stillman, B. (1987) Nature 326, 471–475.

[28] Prelich, G. and Stillman, B. (1988) Cell 53, 117–126.

[29] Celis, J.E. and Madsen, P. (1986) FEBS Lett. 209, 277–283.

[30] Celis, J.E. Bravo, R., Mose Larsen, P. and Fey, S.J. (1984) Leuk. Res. 8, 143–157.

[31] Celis, J.E. and Bravo, R. (1984) in: Electrophoresis 4 (Neuhoff, V. ed.) pp.205–225, Verlag Chemie, Weinheim.

[32] Celis, J.E., Madsen, P., Nielsen, S.U., Gesser, B., Nielsen, H.V., Ratz, G.P., Lauridsen, J. and Celis, A. (1987) FEBS Lett. 220, 1–7.

[33] Bravo, R., Fey, S.J., Bellatin, J., Mose Larsen, P., Arevalo, J. and Celis, J.E. (1981) Exp. Cell Res. 136, 311–319.

[34] Ogata, K., Kurki, P., Celis, J.E., Nakamura, R.M. and Tan, E.M. (1987) Exp. Cell Res. 168, 475–486.

[35] Bravo, R., Frank, R., Blundell, P.A. and Mcdonald-Bravo, H. (1987) Nature 326, 515–517.

[36] Tan, E.M., Ogata, K. and Takasaki, Y. (1987) J. Rheumatol. 14, 89–96.

[37] Tan, C.K., Castillo, C., So, A.G. and Downey, K.M. (1986) J. Biol. Chem. 261, 12310–12316.

[38] Prelich, G., Tan, C.K., Kostura, M., Mathews, M.B., So, A.G., Downey, K.M. and Stillman, B. (1987) Nature 326, 517–520.

[39] Ogata, K., Ogata, Y., Nakamura, R. and Tan, E.M. (1986) J. Immunol. 135, 2623–2627.

[40] Bravo, R. and Mcdonald-Bravo, H. (1985) EMBO J. 4, 655–661.

[41] Bravo, R. and Celis, J.E. (1982) Clin. Chem. 28, 766–781.

[42] Fey, S.J., Bravo, R., Mose Larsen, P. and Celis, J.E. (1984) in: Two-Dimensional Gel Electrophoresis of Proteins: Methods and Applications (Celis, J.E. and Bravo, R. eds) pp.169–189, Academic Press, New York.