Bengt Bjellqvist*
Bodil Basse
Eydfinnur Olsen
Julio E. Celis

Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, Aarhus

Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions

A highly reproducible, commercial and nonlinear, wide-range immobilized pH gradient (IPG) was used to generate two-dimensional (2-D) gel maps of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human epidermal keratinocytes. Forty one proteins, common to most human cell types and recorded in the human keratinocyte 2-D gel protein database were identified in the 2-D gel maps and their isoelectric points (p*I*) were determined using narrow-range IPGs. The latter established a pH scale that allowed comparisons between 2-D gel maps generated either with other IPGs in the first dimension or with different human protein samples. Of the 41 proteins identified, a subset of 18 was defined as suitable to evaluate the correlation between calculated and experimental p*I* values for polypeptides with known composition. The variance calculated for the discrepancies between calculated and experimental p*I* values for these proteins was 0.001 pH units. Comparison of the values by the *t*-test for dependent samples (paired test) gave a *p*-level of 0.49, indicating that there is no significant difference between the calculated and experimental p*I* values. The precision of the calculated values depended on the buffer capacity of the proteins, and on average, it improved with increased buffer capacity. As shown here, the widely available information on protein sequences cannot, *a priori*, be assumed to be sufficient for calculating p*I* values because post-translational modifications, in particular *N*-terminal blockage, pose a major problem. Of the 36 proteins analyzed in this study, 18–20 were found to be *N*-terminally blocked and of these only 6 were indicated as such in databases. The probability of *N*-terminal blockage depended on the nature of the *N*-terminal group. Twenty six of the proteins had either M, S or A as *N*-terminal amino acids and of these 17–19 were blocked. Only 1 in 10 proteins containing other *N*-terminal groups were blocked.

1 Introduction

As compared with carrier ampholyte isoelectric focusing (CA-IEF), the application of immobilized pH gradients (IPGs) in the first dimension in 2-D gel electrophoresis offers improved reproducibility [1] because the nature of the pH gradient makes the resulting focusing positions insensitive to the focusing time [2] and to the type of sample applied [3]. The recently introduced ready-made IPG strips [4] seem to be an ideal substitute for the carrier ampholyte gradients, which until now have been the most commonly used first dimensions in 2-D gel electrophoresis. The availability of standardized first dimensions opens the possibility of comparing 2-D gel maps of various cell types generated in different laboratories, provided that the focusing positions of a number of easily recognizable polypeptide spots common to the cell types in question are known. Even though this approach is limited to experiments performed with the same standardized IPG, the flexibility provided by IPGs allows the pH gradient to be adjusted to the requirements of a particular experiment.

Exchange and communication of 2-D gel protein data requires a pH scale that is independent of the particular IPG used and by which the results can be described. The introduction of carbamylation trains and the relation of focusing positions to the spots in these trains represented a step forward towards solving the reproducibility problem experienced with carrier ampholyte focusing [5]. Problems associated with the use of carbamylation trains were mainly due to lack of temperature control and to the use of nonequilibrium focusing conditions. Accordingly, the pattern variation involved not only the resulting pH gradients, but also the relative spot positions as related to each other and to spots in the carbamylation trains. Even though the question of reproducibility has, to a large extent, been solved, the carbamylation trains are still not ideal as markers because the spots in the trains do not represent defined entities but rather a large number of differently carbamylated peptides having close p*I* values. As a result, the spots are large and poorly defined as compared to the ordinary polypeptide spots in 2-D gel maps.

Correspondence: Professor J. E. Celis, Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, DK-8000 Aarhus C, Denmark

Abbreviations: CA-IEF, carrier ampholyte-isoelectric focusing; **SSP,** sample spot number

* Present addess: Pharmacia Biotech AB, S-751 82 Uppsala, Sweden

0173-0835/94/0304-0529 $5.00+.25/0

Neidhardt *et al.* [6] defined the pH gradient in 2-D gel experiments by p*I* markers whose p*I* values were calculated from the amino acid composition. Focusing positions of other polypeptides could be predicted from their composition but the p*K* values needed for the p*I* calculations were unknown. Various groups employing this approach do not use the same pK values [6, 7] and therefore, the p*I* values derived in this way cannot be expected to describe the variation of the hydrogen ion activity. In spite of this fact, it is still possible to make approximate predictions of focusing positions because the p*K* values used to define the pH gradient are also used to calculate p*I* values and to predict the focusing positions. Errors in p*K* assignments are therefore compensated. A pH scale which corretly reflects the variation in hydrogen ion activity during focusing should improve the precision of the predictions, but this has never been implemented with CA-IEF focusing as a first dimension in 2-D gel electrophoresis. The main reason for this are the problems associated with pH measurements in focused gels containing high concentrations of urea.

IPGs can be described from the concentration variation of the immobilized groups, provided that the p*K* values of these groups are known for the conditions prevailing during focusing. To avoid measurements on gels, Gianazza *et al.* [8] suggested the use of p*K* values derived by addition of determined p*K* shifts. Recently, direct determinations of p*K* differences between immobilized groups in IPGs were made by determining p*I*-p*K* values in overlapping narrow-range IPGs [9, 10] and the results verified the applicability of the Gianazza approach. A description of the focusing results in a pH scale, which correctly describes the variation of the hydrogen ion activity for the focusing conditions used, not only allows the comparison of 2-D gel maps generated with different IPGs, but also opens the possibility for correlating the focusing position of a polypeptide with its composition [9]. Experiments by Bjellqvist *et al.* [9, 10] have implied that pH scales showing good correlation between calculated and experimental p*I* values can be derived for any of the conditions commonly used for focusing in connection with 2-D gel electrophoresis. These pH scales are then defined through the p*K* values of the immobilized groups in the IPG containing gel. To be useful for interlaboratory comparisons, however, the pH scale has to be defined through p*I* values of easily recognizable spots present in the 2-D gel map. So far, p*I* determinations in a useful pH scale, combined with determinations of pK values needed for p*I* calculations, have only been made for the pH range 4.5–6.5 at 10°C [9]. CA-IEF focusing as described by O'Farrell [11] does not control the temperature of the first dimension, which can be expected to be slightly above room temperature. With IPGs, the temperature commonly used is about 20°C [4, 12] or 25°C [13] and this is a critical parameter that needs to be controlled [14].

The present work was designed to compare 2-D gel maps of different cell types in a laboratory applying both CA-IEF and IPG focusing at a common temperature. To this end we have generated 2-D gel maps of proteins from noncultured, unfractionated normal human epidermal keratinocytes with IPG in the first dimension and a focusing temperature of 25°C. We have used commercial nonlinear, wide-range IPG strips which give 2-D gel maps that are closely similar to the ones resulting with the CA-IEF technique used to establish the human keratinocyte database [15]. As an initial step towards interlaboratory comparisons of results obtained with the nonlinear gradient as a first dimension we report here on the focusing positions of 41 known proteins that are common to most human cell types. The pH range covered corresponds to the range in classical CA-IEF 2-D gel electrophoresis and in order to use these proteins as internal standards for comparing 2-D gel maps generated with other IPGs we determined their p*I* values with narrow-range IPGs in the first dimension. We have compared the calculated *versus* experimental p*I* values and show that it is necessary to have further information (absence or presence and nature of posttranslational modifications), in addition to amino acid composition to be able to calculate p*I* values that correspond to the actual experimental values. The p*K* values used for the calculations are provided and the usefulness of p*I* prediction in relation to database information is discussed. Furthermore, we comment on the possibility of using experimentally determined p*I* values to verify the available database information on polypeptide composition.

2 Materials and methods

2.1 Apparatus and chemicals

Equipment for isoelectric focusing and horizontal SDS electrophoresis (Multiphor® II electrophoresis chamber, Immobiline® strip tray, Multidrive XL programmable power supply, Macrodrive power supply and Multitemp® II) was from Pharmacia LKB Biotechnology AB (Uppsala, Sweden). Vertical second-dimensional gels were run in the home-made equipment described in [15]. The IPG strips with the wide-range nonlinear pH gradient were either Immobiline DryStrip® pH 3–10 NL, 180 mm or alternatively 160 mm long IPG strips with a corresponding pH gradient. In both cases the IPG strips were delivered by Pharmacia LKB. Immobiline, Pharmalyte, Ampholine, GelBond as well as PAG film and the ready-made horizontal SDS gels (ExcelGel® XL SDS 12–14) were also from Pharmacia LKB. Purified proteins and peptides were from Sigma (St. Louis, MO).

2.2 Sample preparation

Preparation and labeling of unfractionated keratinocytes as well as fibroblasts have been described in [16]. Cells were lysed in a solution containing 9.8 M urea, 2% w/v NP-40, 100 mM DTT and 2% v/v Ampholine pH 7–9.

2.3 2-D gel electrophoresis

First-dimensional focusing was performed according to Görg *et al.* [2] with some minor modifications, as described in [9]. Rehydration of the IPG strips was made in a solution containing 9.8 M urea, 2% w/v CHAPS, 10 mM DTT and 2% v/v carrier ampholyte mixture. The carrier ampholyte mixture consisted of 2 parts Pharmalyte

4–6.5, 1 part Ampholine pH 6–8 and 1 part Pharmalyte pH 8–10.5. Usually, cathodic sample application was used and the samples were diluted 2–20 times in a solution containing 9.8 M urea, 4% w/v CHAPS, 1% w/v DTT and 35 mM Tris base. For acidic application, the Tris-base was substituted with 100 mM acetic acid. The degree of dilution and sample volume (20–100 µL) depended on the particular sample and the IPG, and whether visualization of the proteins was to be done by Coomassie Brilliant Blue or silver staining. With the wide-range non-linear IPG, 10–30 µg of total protein was loaded for silver staining and 100–200 µg for Coomassie staining. Focusing was done overnight with Vh products in the range of 45–60 kVh with 160 mm long strips and 50–70 kVh with 180 mm long strips. Solubilization of polypeptides and blocking of -SH groups prior to the second-dimensional run, as well as loading on the second-dimensional gel was done as described in [9]. The stacking gel was omitted and 5–10 mm were left at the top of the second-dimensional gel for applying the IPG strip. The space was filled with electrode buffer containing 0.5% w/v agarose. Casting, running, staining and autoradiography were carried out as described in [15].

2.4 Experimental determination of p*I* values

The determination of the p*K* differences between Immobilines p*K* 4.6, p*K* 6.2 and p*K* 7.0 necessary for the calibration of the pH scale at 25°C in 9.8 M urea was done as described in [9] with the same narrow-range IPGs. The pH scale was defined by setting the p*K* value of Immobiline p*K* 4.6 equal to 4.61 [9] and the determined p*K* differences gave the p*K* values of Immobilines p*K* 6.2 and p*K* 7.0, equal to 5.73 and 6.54, respectively. The p*K* differences found are in good agreement with values derived from [17] and [8] by extrapolation to 9.8 M urea concentration. As in [9], additional narrow-range recipes have been used for determining p*I* values. With narrow-range IPGs extending to pH values higher than the p*K* value of Immobiline p*K* 7.0, anodic sample application was used with acetic acid added to the sample solution. Otherwise, cathodic sample application was used with the same sample buffer as for wide-range IPGs.

2.5 Protein compositions used for p*I* calculations

With the exception of vimentin, protein compositions are from the Swiss-Prot database [18]. For vimentin, we used the data from [19], where the amino acid at position 41 is a D instead of a S. Information in the Swiss-Prot database on phosphorylation has been disregarded because it was known from earlier studies (J. E. Celis, unpublished results) that the spots in question corresponded to the unphosphorylated forms of the peptides.

2.6 Calculation of p*I* values

For the p*I* calculations it was assumed that the same p*K* value could be used for an amino acid residue in all polypeptides and in all positions in the peptide except for *N*- or *C*-terminally placed amino acids. For the p*K* values of the *N*-terminal amino groups the effect of the different substituents on the α-carbon were taken into account. The calculations of p*I* values were made with the aid of the IPG-maker program [20].

2.7 p*K* values used for p*I* calculations

For the carboxyl terminal group and internal glutamyl and aspartyl residues the same p*K* values were used as in [9]. For *C*-terminal glutamyl and aspartyl residues, separate p*K* values were derived with the aid of the Taft equations [9, 21]. The p*K* values of histidyl groups were calculated from the p*I* values of human carbonic anhydrase I as in [9]. For *N*-terminal glycine a p*K* value of 7.50 was used. The p*K* shift caused by a substituent on the α-carbon was assumed to be identical with the p*K* shift the substituent caused for the amino group in the amino acid, *i.e.* 2.28 pH units were subtracted from the p*K* values for the amino groups in the amino acids given in [22, 23]. The approximate p*K* value of 9 for the cystenyl group was taken from [24]. For tyrosyl and arginyl groups we used the p*K* values for the amino acids [22, 23]. For lysyl groups the effect of high urea concentration on amino groups was taken into account and 0.5 pH units were subtracted from the amino acid p*K* value. These last three p*K* values are far from the pH range under study and the results found would have been the same if lysyl and arginyl groups were assumed to be fully ionized while the ionization of tyrosyl groups were neglected. A complete list of the p*K* values used is given in Table 1.

Table 1. p*K* Values used for the ionizable groups in peptides 9.8 M urea, 25°C

Ionizable group	p*K*
C-terminal	3.55
N-terminal	
Ala	7.59
Met	7.00
Ser	6.93
Pro	8.36
Thr	6.82
Val	7.44
Glu	7.70
Internal	
Asp	4.05
Glu	4.45
His	5.98
Cys	9
Tyr	10
Lys	10
Arg	12
C-terminal side chain groups	
Asp	4.55
Glu	4.75

2.8 Statistical analysis

Statistical comparisons of the experimental and calculated p*I* values were done on an Apple Macintosh IIsi using the statistical package Statistica/Mac, release 3.0b (from StatSoft Inc., Tulsa, Oklahoma). Calculated and experimental p*I* values were compared by the *t*-test for correlated samples (paired *t*-test). The normality of p*I* differences was estimated graphically by probability plots. The variances of the data presented here and the similar data on plasma and liver proteins in [9] were compared by the F-test.

3 Results and discussion

3.1 Identification of polypeptides and p*I* determinations

The 2-D gel maps of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes, focused with the nonlinear, wide-range IPG and CA-IEF pH gradients in the first dimension, are shown in Figs. 1 and 2, respectively. The IPG extends to higher pH values but otherwise the two patterns are very similar and most of the spots in the IPG pattern can be directly related to the corresponding spots in the CA-IEF gel. To obtain comparable patterns it was important to keep the focusing temperature as similar as possible. Compared to other studies [1–4, 9, 10, 12–14], we increased the urea concentration in the focusing gel to 9.8 M because keratins streaked badly in the focusing dimension when 8 M urea was used, presumably due to

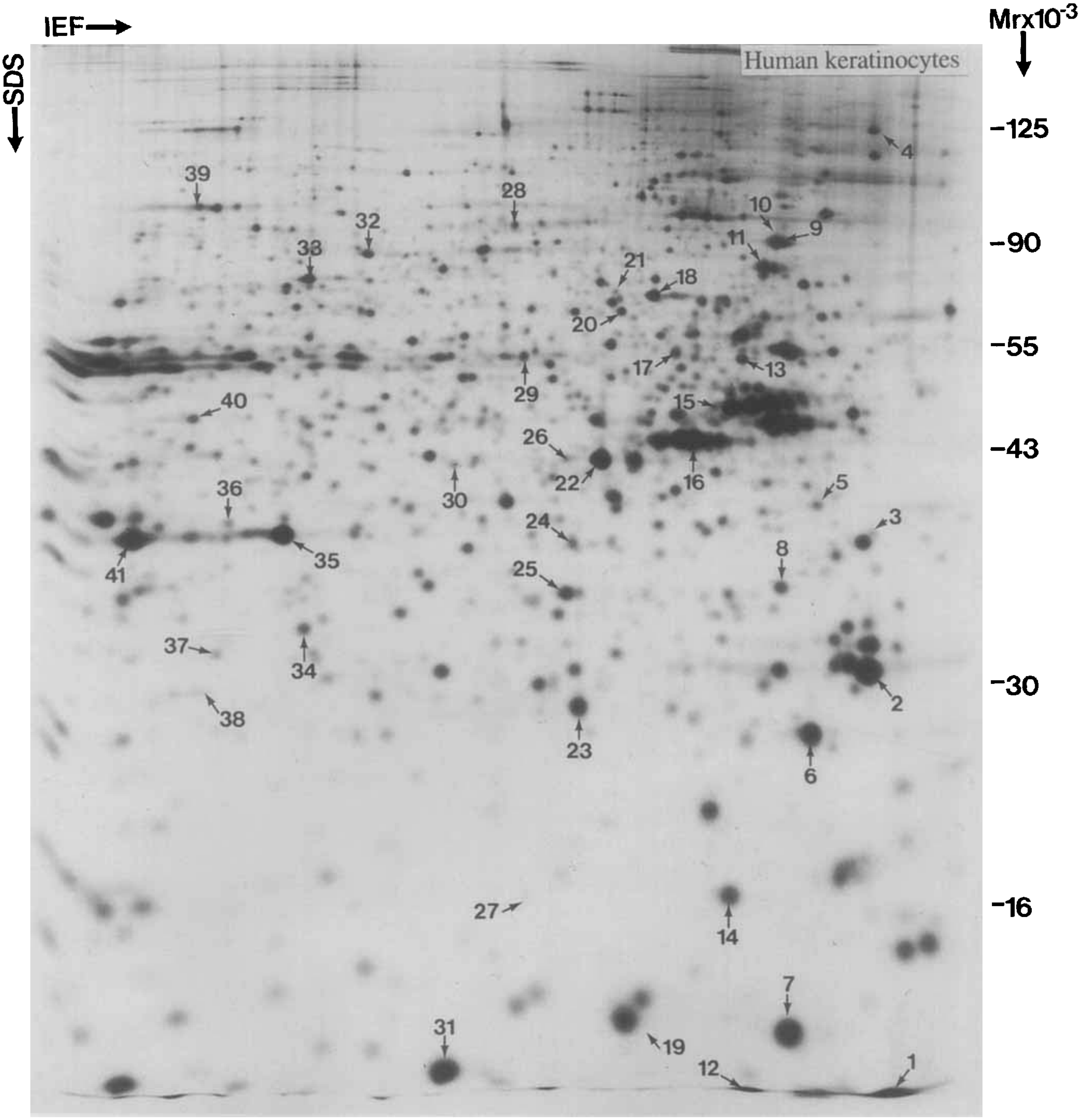

Figure 1. 2-D gel protein map of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

aggregates of acidic and basic keratins. An increase in urea concentration to 9 M or more eliminated these streaks; apart from this effect, no other major changes in the focusing positions were observed. In Fig. 1 we have indicated the positions of 41 known proteins from the human keratinocyte 2-D gel database that are most likely common to most human cell types. The choice was made because these proteins are easy to identify with certainty. With the exception of stratifin (spot 2), involucrin (spot 4) and keratin 14 (spot 15), which are all epithelial markers, these proteins are also present in human fibroblasts (Fig. 3) and lymphocytes (results not shown), and therefore can be used as landmarks for comparing 2-D gel maps derived from different cell types. In Table 2 the 41 proteins are listed together with their sample spot numbers (SSP) in the human keratinocyte protein database and p*I* values determined in 2-D gel maps generated with narrow-range IPGs in the first dimension.

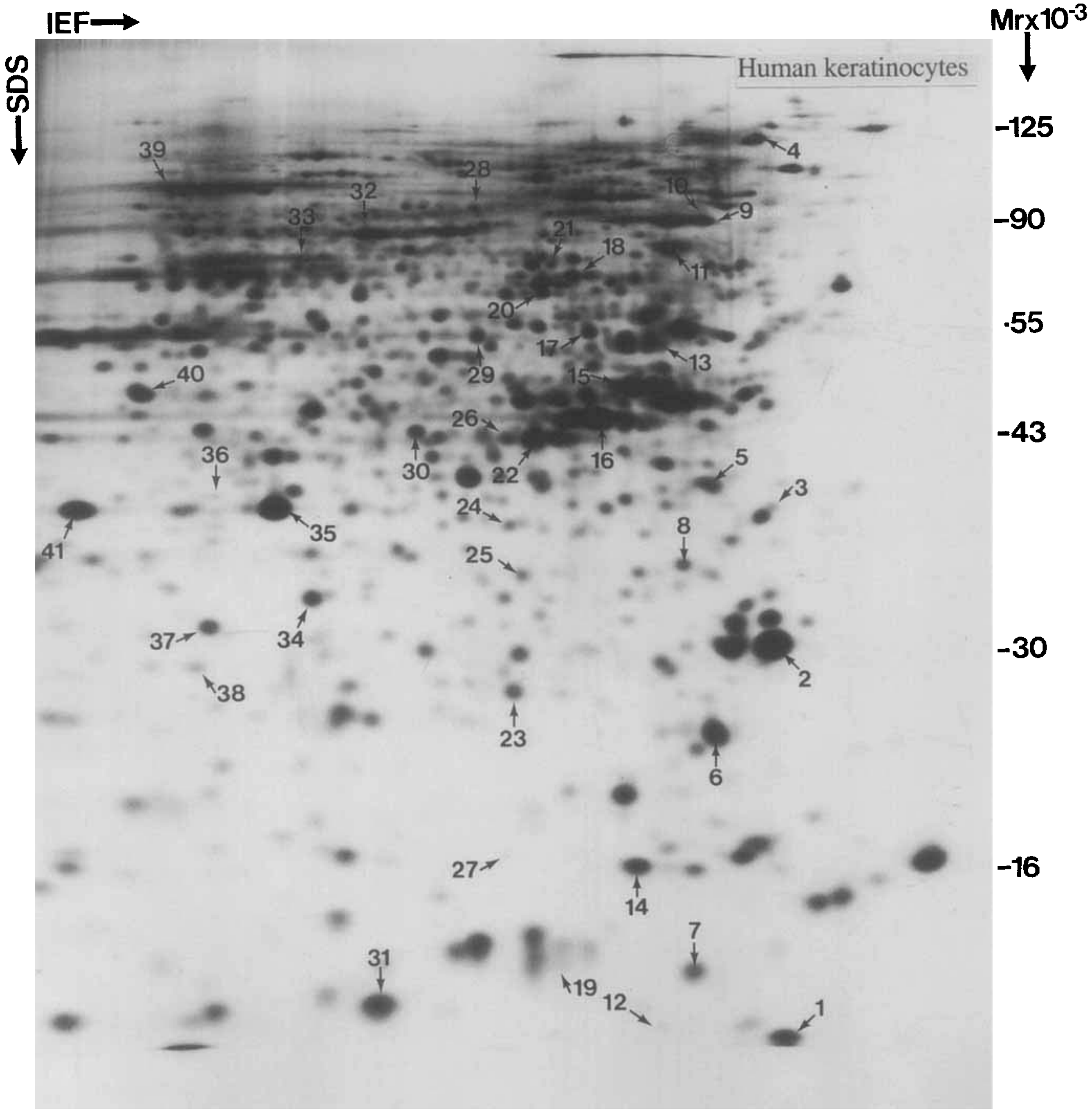

Figure 2. 2-D gel protein map of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with CA-IEF in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

Table 2. Proteins from the human keratinocyte database localized in 2-D gels run with IPGs as first dimension

Number in Figs. 1–3	Protein name	IEF SSP number[a]	Experimental p*I* value	Calculated p*I* value	Discrepancy (pH units)	Calculated net charge at experimental p*I* value	Buffer capacity charge units pro pH unit	*N*-terminal	Recalculated for suspected blockage		*N*-terminal	Swiss-Prot accession number
									p*I* value	Discrepancy pH units	Net charge	
1	CaN 19	9027	4.46	–	–	–	–	–				–
2	Stratifin, bovine 14-3-3 related protein	9109	4.58	–	–	–	–	–				–
3	Proliferating nuclear antigen (PCNA)/cyclin	9226	4.58	4.57	–0.01	–0.1	20.8	M				P12004
4	Involucrin	9703	4.63	4.63	0.00	–0.3	70.1	M				P07476
5	Nucleolar protein B23	8207	4.75	4.64	–0.11	–3.2	30.4	M				P06748
6	Translationally controlled tumor protein	8114	4.79	4.84	0.05	0.6	13.1	M[b]				P13693
7	Thioredoxin	8006	4.86	4.82	–0.04	–0.3	7.1	V[b]				P10599
8	Annexin V	8213	4.89	4.88	–0.01	–0.1	20.3	A[c]				P08758
9	Heat shock protein 90-β	8611	4.95	4.94	–0.01	–0.5	56.2	P				P07900
10	Heat shock protein 90-α	2629	4.97	4.97	0.00	–0.2	53.6	P				P08238
11	Glucose regulated protein 78 (BiP)	8515	4.99	4.98	–0.01	–0.6	37.5	E				P11021
12	Calcyclin	8017	5.02	5.32	0.30	1.3	3.6	M	5.09	0.07	0.3	P06703
13	Vimentin	8417	5.05	5.06	0.01	0.2	27.1	S				P08670
14	Initiation factor 4D	8016	5.05	5.08	0.03	0.2	7.6	A[c]				P10159
15	Keratin 14	7305	5.08	5.09	0.01	0.2	21.0	T				P02533
16	β-Actin	7316	5.21	5.21	0.00	0.06	13.3	D[c]				P02570
17	Heat shock protein 60	6403	5.23	5.24	0.01	0.1	17.5	A[b]				P10809
18	Heat shock cognate 71kD	6504	5.28	5.37	0.09	1.8	18.1	M	5.32	0.04	0.8	P11142
19	Cystatin	6011	5.30	5.38	0.08	0.2	3.0	M				P01040
20	T-plastin	6412	5.34	5.41	0.07	1.3	17.7	M	5.36	0.02	0.3	P13797
21	Calelectrin	5628	5.35	5.37	0.02	0.5	23.3	A[c]				P08133
22	Plasminogen activator inhibitor-2	6314	5.38	5.46	0.08	0.9	10.7	M	5.37	–0.01	–0.07	P05120
23	Glutathione S-transferase π	5101	5.43	5.44	0.01	0.08	3.9	P				P09211
24	Annexin VIII	5213	5.45	5.56	0.11	1.0	8.7	M	5.46	0.01	0.05	P13928
25	Annexin III	5204	5.46	5.63	0.17	1.4	8.4	M	5.52	0.06	0.5	P12429
26	Adenosine deaminase	5305	5.47	5.63	0.16	1.8	10.8	M	5.54	0.07	0.8	P00813
27	Stathmin	5001	5.55	5.61	0.06	0.4	6.6	A[c]				P16949
28	Gelsolin, cytoplasmic	5608	5.59	5.58	–0.01	–0.1	16.5	V				P06396
29	Rat phosphoinoside specific protein homolog	5410	5.62	–	–	–	–	–				–
30	Elastase inhibitor	4314	5.74	–	–	–	–	–				–
31	S100, calgizarin	4006	5.75	–	–	–	–	–				–
32	Cytvillin, ezrin	3504	5.99	5.95	–0.04	–0.5	13.2	P				P15311
33	Moesin	3515	6.11	6.09	–0.02	–0.2	9.8	P				P26038
34	Purine nucleoside phosphorylase	2108	6.11	6.45	0.34	1.8	4.4	M	6.28	0.17	0.9	P00491
35	Annexin I	2216	6.18	6.64	0.46	1.6	2.5	A	6.33	0.15	0.6	P04083
36	Aldose reductase	1202	6.40	6.55	0.15	0.7	4.2	A	6.36	–0.04	–0.2	P15121
37	Phosphoglycerate mutase (B form)	1107	6.46	6.75	0.29	0.9	2.6	A	6.46	0.00	0.0	P18669
38	Triosephosphate isomerase	1111	6.53	6.51	–0.02	–0.04	2.3	A[b]				P00938
39	Elongation factor 2	1610	6.43	6.38	–0.05	–0.5	9.8	M				P13639
40	α-Enolase	1325	6.62	6.99	0.37	1.0	2.2	S	6.75	0.13	0.3	P06733
41	Annexin II	210	7.30	7.36	0.06	0.05	0.9	S[c]				P07355

a) SSP number in the keratinocyte database [15]

b) Peptides *N*-terminally sequenced as liver proteins [3]

c) Peptides given as *N*-terminally blocked in Swiss-Prot database

3.2 Comparison between the determined and calculated p*I* values for human keratinocyte proteins

Thirty six of the 41 proteins listed in Table 2 are found in the Swiss-Prot database. Contrary to the plasma and liver proteins used in [9], the p*I* calcuations on the proteins used in this study posed some problems that reflected the way in which they were characterized. The proteins used by Bjellqvist *et al.* [9] were either very abundant and well-characterized plasma proteins or they were identified by *N*-terminal sequencing and, therefore, the nature of the *N*-terminals (acetylated or non-acetylated) was in both cases known. The proteins used in this study have all been characterized by internal sequencing [7] and it is known that *N*-terminal acetylation occurs with high frequency in eukaryotes.

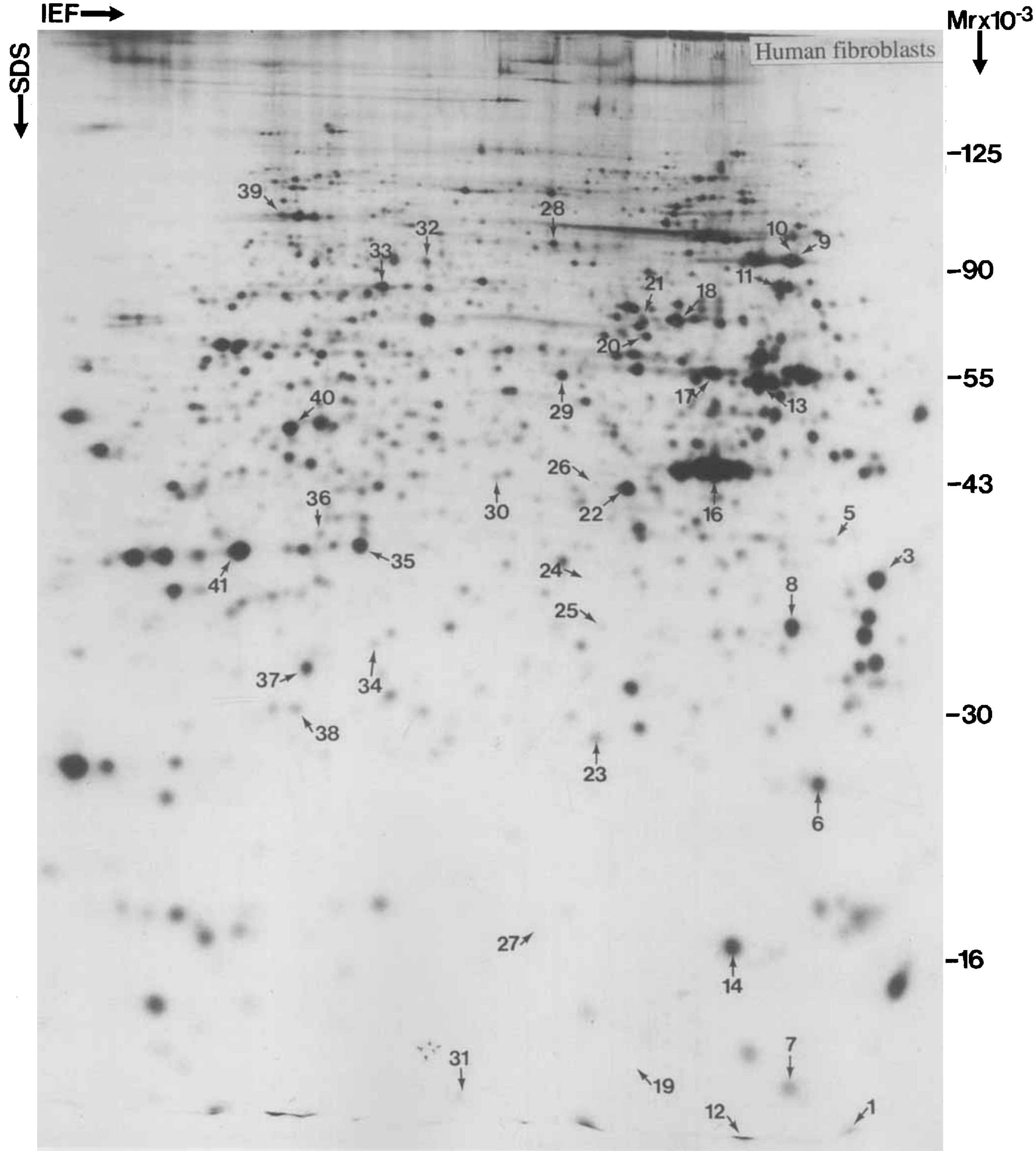

Figure 3. 2-D protein map of [^{35}S]methionine-labeled proteins from normal human fibroblasts focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

According to Brown and Robert [25], proteins with acetylated *N*-terminals correspond in weight to approximately 80% of the soluble protein in ascites cells. Based on results from *N*-terminal sequencing, at least 40% of the spots in the human liver protein 2-D gel map appear to be blocked [3]. The corresponding number, derived from 107 spots in the 2-D gel map of human T-lymphocyte proteins, falls between 60 and 65% (J. Strahler, personal communication). Information concerning *N*-terminal blockage is not normally available, and in the Swiss-Prot database only 6 of the 36 keratinocyte proteins are specified as *N*-terminally blocked. We have, within the present material, defined 18 proteins for which the *N*-terminals are very likely to be correctly described. Six of these proteins are listed in the Swiss-Prot database as *N*-terminally blocked, four represent proteins which appear in the human liver 2-D gel map and have been *N*-terminally sequenced as liver proteins [3] and the remaining eight have *N*-terminal groups other than M, S and A, *i.e.* *N*-terminals for which *N*-acetylation is uncommon [26]. In Figs. 4A, B, C and D p*I* values calculated from Swiss Prot database information are plotted against the experimentally determined p*I* values for all the keratinocyte proteins listed in Table 2 and for the 18 selected proteins, as well as for the plasma and liver proteins (data from [9] valid for 10°C)*.

The calculations show that without knowledge of the status of the *N*-terminal group, precise predictions of p*I* values for eukaryotic proteins cannot be achieved based on the information available in Swiss-Prot and similar databases. However, for proteins where the *N*-terminal status is known, we find good correlation between predicted and experimental p*I* values. When the variance of the p*I* discrepancies and the variance of calculated charges at the experimental p*I* values derived from the present data set are compared with the corresponding

* There are four plots: (A) the 36 polypeptides from normal human keratinocytes (no corrections), (B) the 36 polypeptides from Fig. 4A where p*I* values have been recalculated for 12 polypeptides with M, S and A as *N*-terminally assumed blocked, based on calculated charge, (C) the 18 selected polypeptides with information on the *N*-terminal configuration, and (D) plasma and liver proteins.

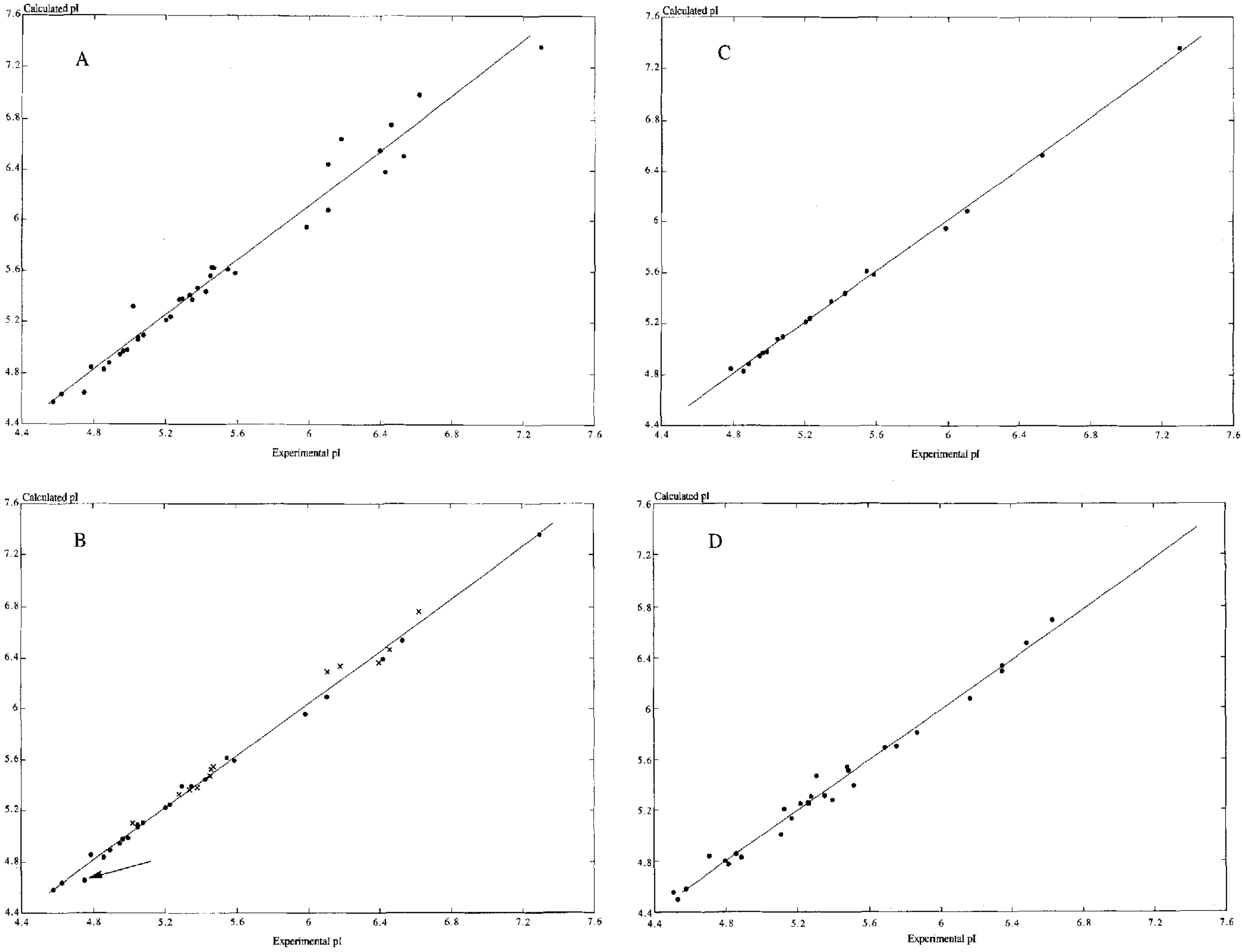

Figure 4. Calculated *vs.* experimental p*I* values. Lines are fitted using the least squares' criterion. (A) 36 polypeptides from normal human keratinocytes (no corrections). (B) 36 polypeptides from Fig. 4A (including the 18 marker polypeptides) where p*I* values have been recalculated assuming *N*-terminal blockage; x indicates recalculated p*I* values; nucleolar protein B23 is indicated with an arrow. (C) 18 polypeptides with information on *N*-terminal configuration and (D) plasma and liver proteins.

values derived from the data on plasma and liver proteins in [9] (Table 3), the present data are found to result in larger variances for the values of both p*I* discrepancies and calculated charge at the experimental p*I* value when no information on posttranslational modification is taken into consideration. Correction for possible *N*-acetylation of 12 polypeptides with M, S and A as *N*-terminal results in a smaller variance of p*I* discrepancies, although not significantly different from values derived from [9], whereas the variance of the calculated charge at the experimental p*I* value is significantly higher. For the 18 selected proteins the variance for the p*I* discrepancies is significantly smaller than for the data in [9]; however, the corresponding value for calculated charge at the experimental p*I* value does not improve to the same extent. This, we believe, reflects another difference between the two sets of proteins used for the calculations. Based on spot distributions in 2-D gel maps, the set of proteins used here has a molecular weight distribution that is more representative of the patterns observed in mammalian cells. In the study by Bjellqvist *et al.* [9] most of the high molecular weight plasma proteins had to be excluded due to their unknown content of sialic acid which made the proteins analyzed in this study heavily biased towards low molecular weight proteins. The buffer capacity of proteins normally increases with the protein's molecular weight, and the average buffer capacity of the presently selected proteins with assumed known *N*-terminals is 18 charge units/pH unit, while the corresponding value for the proteins used in [9] is only 9 charge units/pH unit. High buffer capacity can be expected to improve the agreement between calculated and experimental p*I* values. Inspection of the data presented in Table 2 for the polypeptides with assumed known *N*-terminals verifies the importance of the buffer capacity. For 8 polypeptides having buffer capacities higher than 15 charge units/pH unit, the calculations in all cases yielded p*I* discrepancies with absolute values of less than 0.02 pH units. The largest discrepancy, 0.06 pH units, was observed for annexin II and stathmin, proteins which have low buffer capacity: 0.9 and 6.6 charge units/pH unit, respectively. The probability that the focusing position of a protein with known composition will fall within a certain distance from the calculated p*I* value therefore cannot be predicted by the variance alone. The buffer capacity of the specific protein must be taken into consideration as well. As indicated by the decrease of the variance of calculated charges at the experimental p*I* value for the selected proteins, the observed improvement can not solely be due to the higher buffer capacity of the keratinocyte proteins. The two studies relate to different experimental conditions. Good agreement between experimental and calculated p*I* values implies that the proteins are defolded and a factor that may contribute to the observed improvement is a more complete defolding of proteins caused by the higher temperature and urea concentration used in this study.

The data indicated that the precision with which p*I* values can be predicted for polypeptides with high buffer capacity is better than the precision with which experimental p*I* values can be determined. If the pH is defined through the p*K* values of the immobilized groups in the IPG containing gel, the precision of the experimentally calculated data will depend on the pH difference between the p*I* and the p*K* value of the immobilized group with the closest p*K*. For the present study this will give p*I* determinations with a precision varying in the range of ± 0.02–0.05 pH units [9]. The good agreement observed between the calculated and experimental p*I* values is due to the fact that errors are mainly systematic and, as discussed in [9], they will largely be cancelled out in the calculations. A pH scale defined through the presently determined p*I* values will not necessarily reflect the variation of the hydrogen ion activity during the focusing step in an optimal way, but it still allows precise predictions of focusing positions for polypeptides with known compositions, including information on posttranslational modifications. Calculated net charge at the experimentally found isoelectric point defined in this scale will serve as a tool to verify that the polypeptide

Table 3. Mean values and variances for the difference (experimental p*I*-calculated p*I*) in pH units and calculated charges at the experimental p*I* values, respectively

	Plasma and liver proteins (8 M urea, 10°C)		Keratinocyte proteins (9.8 M urea, 25°C)					
			All peptides		All peptides after correction for *N*-acetylation		Known *N*-terminal configuration (or very likely configuration)	
Number of proteins	29		36		36		18	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Experimental p*I*-calculated p*I*	−0.011	0.005	0.072	0.017	0.019	0.003	0.005	0.001
F-value (p*I* discrepancy)[a]	1		3.4		1.67		5	
P-level (p*I* discrepancy)[b]	0.5		0.0005		0.0721		0.0004	
Calculated charge at the experimental p*I* value	−0.070	0.227	0.321	0.871	0.009	0.444	−0.014	0.109
F-value (calculated charge at the experimental p*I* value)[a]	1		3.8		1.96		2.08	
P-level (calculated charge at the experimental p*I* value)[b]	0.5		0.0002		0.0338		0.0536	

a) Comparison to the data in [9]. $F = S_1^2/S_2^2$, where S_1^2 is the larger of the two variances

b) $P(F(\nu_1, \nu_2) \geq F\text{-value})$, where ν_1 and ν_2 are the degrees of freedom for s_1 and s_2, respectively composition used in the calculation is correct and complete. Exceptions to this are proteins such as involucrin and heat shock protein 90 that have very high buffer capacities. Introduction of an extra charge unit into these proteins will only result in p*I* shifts falling in the range of 0.01–0.02 pH units and the effect is that the quality of the pH definition – the precision by which p*K* values used in the calculations are given and the precision of experimental p*I* values in these cases – will limit the possibilities to verify polypeptide compostion based on the experimental p*I* value.

Statistical comparison of experimental and calculated p*I* values was done using the *t*-test for dependent samples and normality of the discrepancies was estimated by probability plots. For the 36 proteins, the *p*-level is 0.0021, indicating that a result like this is unlikely to be a chance effect and must be assumed to represent a real difference. After correction for the most likely *N*-terminal configuration, the *p*-level is 0.043 and cannot be accepted as representing the same population since the *p*-level is less than 0.05 – the traditional *p*-limit of statistical significance. For the 18 proteins with a known or very likely *N*-terminal configuration the *t*-test gave a *p*-level of 0.49, which verifies that the experimental and calculated p*I* values are not significantly different.

Besides showing that p*I* values for denatured proteins with known compositions can be calculated with a high degree of precision from average p*K* values, the results also provide strong support for the notion that *N*-terminal blockage heavily depends on the nature of the *N*-terminal groups [26]. The results seem to indicate that with *N*-terminals other than M, S and A, only a few proteins have blocked *N*-terminals (1 out of 10 proteins in the present study), while it can be inferred from the data presented in Table 2 that a majority of the proteins with M, S and A as *N*-terminal are blocked. After correction for the effect of suspected *N*-terminal blockage there is only one protein (nucleolar protein B23) out of the 36 used in this study, which, in spite of a high buffer capacity, has a marked difference of 0.11 pH units between predicted and determined p*I* values (Fig. 4B); this corresponds to 3 charge units due to the high buffer capacity of this protein. This discrepancy in p*I* prediction and calculation of net charge at the p*I* is probably not due to deficiencies in the database information but instead reflects a shortcoming of the model used for p*I* calculations. Nucleolar protein B23 contains a domain extremely rich in aspartic and glutamic acid residues (Table 4), in which 26 out of 28 amino acid residues from position 161 to 188 are either a D or an E. A calculation based on the use of average p*K* values uninfluenced by the charged neighboring amino acid residues cannot be expected to correctly describe the p*I* value with almost half of the acidic groups packed together into a highly negatively charged region. This limitation caused by calculations based on average p*K* values does not severely limit the usefulness of the approach since a search through Swiss-Prot shows that this type of D/E-rich motif is uncommon, and the existence of a highly charged region is immediately apparent upon inspection of the amino acid sequence.

Table 4. Amino acid sequence of nucleolar phosphoprotein B23

1	MEDSMDMDMS	PLRPQNYLFG	CELKADKDYH	FKVDNDENEH	QLSLRTVSLG
51	AGAKDELHIV	EAEAMNYEGS	PIKVTLATLK	MSVQPTVSLG	GFEITPPVVL
101	RLKCGSGPVH	ISGQHLVAVE	EDAESEDEEE	EDVKLLSISG	KRSAPGGGSK
151	VPQKKVKLAA	DEDDDDDDEE	DDDEDDDDDD	FDDEEAEEKA	PVKKSIRDTP
201	AKNAQKSNQN	GKDSKPSSTP	RSKGQESFKK	QEKTPKTPKG	PSSVEDIKAK
251	MQASIEKGGS	LPKVEAKFIN	YVKNCFRMTD	QEAIQDLWQW	RKSL

The quality of the information available in databases, especially concerning posttranslational modifications, is a major problem when the data is to be used for p*I* predictions. The *p*-level of 0.043 found for all 36 proteins after correction for *N*-acetylation, shows that this problem is not only limited to *N*-terminal blockage and the very good agreement found for the eighteen polypeptides, with assumingly correctly described *N*-terminal (Fig. 4C), must be regarded as an exception from this point of view. *N*-Terminal blockage is generally the main problem in relation to p*I* predictions for eukaryotic proteins. Of the 36 keratinocyte proteins analyzed, 18–20 are suspected to be *N*-terminally blocked (6 proteins blocked according to Swiss-Prot, 12 proteins with M, S or A as *N*-terminal and assumingly blocked based on the calculated charge, and two proteins, involucrin and nucleolar protein B23, with M as *N*-terminal for which the data does not allow any conclusion). This is in reasonable agreement with the conclusions based on the *N*-terminal sequencing data derived in connection with 2-D gel electrophoresis. *N*-terminal blockage can be suspected for 17–19 of the 26 proteins with M, S or A as *N*-terminal, while only 1 in 10 proteins with other *N*-terminal groups are blocked. The information that the frequency of *N*-terminal blockage is strongly related to the nature of the *N*-terminal group will be of some help in connection with p*I* predictions based on database information. However, without information from other sources, an uncertainty will always remain as to whether the *N*-terminal charge should be included in the p*I* calculation.

4 Concluding remarks

The data presented here lays the foundation for comparing 2-D gel protein maps of different cell types generated with nonlinear, wide-range IPGs in the first dimension. The focusing positions of 41 polypeptides common to most human cell types have been described in a pH scale that allows focusing positions to be predicted with a high degree of accuracy, provided that the composition of the polypeptides are known and that information on posttranslational modifications are available. For polypeptides with a very high buffer capacity, the limiting factor is the precision with which experimental pH values can be determined rather than the precision of the calculations. Possible deficiencies in the pH scale description of the variation of the hydrogen ion activity has, at least at the present state, no consequences for its practical use. The major limitation in connection with predictions of focusing positions from polypeptide compositions is the quality of existing data on protein compositions, especially concerning posttranslational modifications. Amino acid sequences have been reasonably easy to obtain, while posttranslational modifications have been difficult and work-intensive to determine. Recent developments in the field of mass spectrometry are fast changing this situation and within the next years we can expect a surge in reliable data in this area. While awaiting this development, verification of correctness and completeness of available information on polypeptide composition can be provided by experimental p*I* values in a pH scale based on the p*I* values determined in this study. So far, our data cover the pH range below pH ≈ 7.5. The basic pH range covered by NEPHGE as first dimension will be covered in forthcoming work.

Received December 29, 1993

5 References

[1] Gianazza, E., Astrua-Testori, S., Caccia, P., Giacon, P., Quaglia, L., Righetti, P. G., *Electrophoresis* 1986, *7*, 76–83.

[2] Görg, A., Postel, W., Günther, S., *Electrophoresis* 1988, *9*, 531–546.

[3] Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Bjellqvist, B., Vargas, R., Appel, R. D., Hughes, G. J., *Electrophoresis* 1992, *13*, 992–1001.

[4] *Immobiline DryStrip Kit for 2-D Electrophoresis: Instructions*, Pharmacia LKB Biotechnology AB, Uppsala 1993.

[5] Anderson, N. L., Hickman, B. J., *Anal. Biochem.* 1979, *93*, 312–320.

[6] Neidhardt, F. C., Appleby, D. A., Sankar, P., Hutton, M. E., Phillips, T. A., *Electrophoresis* 1989, *10*, 116–121.

[7] Rasmussen, H. H., Damme, J. V., Puype, M., Gesser, B., Celis, J. E., Vandekerckhove, J., *Electrophoresis* 1992, *13*, 960–969.

[8] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, *4*, 321–326.

[9] Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1023–1031.

[10] Bjellqvist, B., Pasquali, C., Ravier, C., Sanchez, J.-C., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1357–1365.

[11] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007–4021.

[12] Görg, A., *Biochem. Soc. Transactions* 1993, *21*, 130–132.

[13] Hanash, S. M., Strahler, J. R., Neel, J. V., Hailat, N., Malhem, R., Keim, D., Zhu, X. X., Wagner, D., Gage, D. A., Watson, J. T., *Proc. Natl. Acad. Sci. USA* 1991, *88*, 5709–5713.

[14] Görg, A., Postel, W., Friedrich, C., Kuick, R., Strahler, J. R., Hanash, S. M., *Electrophoresis* 1991, *12*, 653–658.

[15] Celis, J. E., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honoré, B., Dejgaard, K., Gromov, P., Hoffmann, H. J., Nielsen, M., Vassilev, A., Vintermyr, O., Hao, J., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A. H., Walbum, E., Kjærgaard, I., Puype, M., Van Damme, J., Delay, B., Vandekerckhove, J., *Electrophoresis* 1993, *14*, 1091–1198.

[16] 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., Walbum, E., Brandstrup, B., Pedersen, P. S., Brandt, N. J., Puype, M., Van Damme, J., Vandekerckhove, J., *Electrophoresis* 1991, *11*, 802–872.

[17] Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Postel, W., Westermeier, R., *J. Biochem. Biophys. Methods* 1982, *6*, 317–333.

[18] Bairoch, A., Boeckman, B., *Nucleic Acids Res.* 1991, *19*, 2247–2249.

[19] Honoré, B., Madsen, P., Basse, B., Andersen, A., Walbum, E., Celis, J. E., Leffers, H., *Nucleic Acids Res.* 1990, *18*, 6692.

[20] Altland, K., *Electrophoresis* 1990, *11*, 140–147.

[21] Perrin, D. D., Dempsey, B., Serjant, E. P., *pKa Predictions for Organic Acids and Bases*, Chapman and Hall Ltd., London 1981.

[22] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueos Solutions*, Butterworths, London 1965.

[23] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Supplement 1972, Butterworths, London 1972.

[24] Altland, K., Becher, P., Rossman, U., Bjellqvist, B., *Electrophoresis* 1988, *9*, 474–485.

[25] Brown, J. L., Robert, W. K., *J. Biol. Chem.* 1976, *251*, 1009–1014.

[26] Persson, B., Flinta, C., Heine, G., Jörnvall, H., *Eur. J. Biochem.* 1985, *152*, 523–527.