Djamel Medjahed[1,5]
Gary W. Smythers[2,5]
Douglas A. Powell[3]
Robert M. Stephens[2,5]
Peter F. Lemkin[4]
David J. Munroe[1,5]

[1]Laboratory of Molecular Technology
[2]Advanced Biomedical Computing Center
[3]Data Management Services
[4]Laboratory of Computational and Experimental Biology, CCR
[5]Scientific Applications International Corporation Frederick, National Cancer Institute at Frederick, Frederick, MD, USA

VIRTUAL2D: A web-accessible predictive database for proteomics analysis

The available archive of sequence databases compiled from whole genome projects and budding proteomics efforts have enabled us to develop VIRTUAL2D, an interactive system for the assembly of virtual protein expression maps computed on the basis of theoretical isoelectric focusing point, molecular weight, tissue specificity and relative abundance for any set of proteins currently catalogued. This tool will assist in the preliminary, albeit putative, prediction of the identity and location of unknown and/or low abundance proteins in experimentally derived two-dimensional polyacrylamide gel electrophoresis maps.

Keywords: Expressed sequence tags / Protein expression map / Two-dimensional gel electrophoresis / VIRTUAL2D PRO 0340

1 Introduction

One of the most attractive features of 2-D PAGE [1, 2] is its potential to simultaneously display the electrophoretic signatures of complex mixtures of proteins directly from cellular extracts. Continuous progress made over the past two decades in sample treatment, labeling chemistry [3], automation [4], spot detection [5] and image analysis [6] have combined to transform 2-D PAGE from a labor-intensive, multi-process technique to a powerful, highly reproducible tool that is becoming an integral part of many comprehensive proteomic efforts. When used in tandem with other methodologies such as in-gel digestion [7], high resolution mass spectrometry [8] and peptide mapping [9] (many of which can now be run continuously in stand-alone mode), it can provide the front-end for a high-throughput peptide identification scheme [10]. The utility of such schemes are represented by several richly annotated and comprehensive proteomics databases many of which are publicly available on the World Wide Web [11]. Unfortunately, these techniques require substantial financial, labor and time commitments, which place them out of the scope of many laboratories. In addition, because these are developed and maintained by teams of researchers focusing on a particular class of proteins or diseases, typically only a very small fraction of the expressed proteins are identified on these protein expression maps (PEM) [12]. Thus evolved the need to develop a convenient and instantaneous means to predictably identify the unknown protein spots, which in fact constitute the overwhelming proportion of 2-D PAGE data currently available.

Correspondence: Dr. Djamel Medjahed, National Cancer Institute at Frederick, P.O. Box B, Frederick, Maryland 21702-1201, USA
E-mail: medjahed@ncifcrf.gov
Fax: +1-301-846-6827

2 Materials and methods

2.1 Modeling and computational approach

The advent of immobilized pH gradients [13] in the first dimension has ushered in an era where reproducible, high-resolution measurements can routinely be carried out, making it conceivable to predict from the primary sequence, the focusing positions of proteins within a pH gradient. When solubilized with high concentrations of urea (9–10 M), proteins unfold and only the ionizable groups or those amino acids located at the *N*- or *C*-terminally amino acids will affect the electrophoretic mobility of the extended conformation. Using a series of well characterized peptides, Bjellqvist [14] determined the pK values of all the amino acids in similar experimental conditions. The approach that we used to determine the isoelectric focusing point and molecular mass of a peptide can then simply be summed up as follows: (i) scan the primary sequence of the peptide, (ii) assign the pK of each contributing amino acid according to Table 1, (iii) sum up all the mass contributions.

The resulting p*I*/M_r is then given by the ratio of:

$$pI_{tot} = \frac{\{pK_{Cterm} + \Sigma_{int} pK_{int} + pK_{Nterm}\}}{(n+2)} \quad (1)$$

 0173-0835/03/0202–129 $17.50+.50/0

Table 1. Values of amino acid masses and pK's (determined [13] at high molar concentrations of urea) used in p*I*//M_r computation. The segregation is underscored by the fact that the pK's of roughly half the internal amino acids fall below pH 6.0 while for the rest they are greater than or equal to 9.0.

Ionizable group	pK	Molecular Mass
C-terminal	3.55	
N-terminal		
Met	7.00	132.994
Thr	6.82	102.907
Ser	6.93	88.8800
Ala	7.59	72.8800
Val	7.44	100.934
Gli	7.70	130.917
Pro	8.36	98.9180
Internal		
Asp	4.05	116.890
Glu	4.45	130.917
His	5.98	138.943
Cys	9.00	104.940
Tyr	10.0	164.978
Lys	10.0	114.961
Arg	12.0	157.989
C-terminal side chain groups		
Asp	4.55	116.890
Glu	4.75	130.917

and

$M_{r\,tot} = \Sigma_i\, M_r i$

Where the p*I* summation runs over all *n* contributing, *internal* amino-acids.

2.2 Database mining

Queries using the Sequence Retrieval System (SRS) [15] were carried out for each organism against the latest combined releases of SWISS-PROT and TrEMBL databases [16]. Using additional Boolean logic conditions, peptide fragments were omitted from the final p*I*//M_r computation. ExPASy's p*I*//M_r tool [17] server was initially used. To overcome data throughput bottlenecks that are inherent to networks and web browsers, Perl scripts have been developed to locally process files containing individual or multiple sequences in FASTA format. These could represent entire proteomes and ultimately whole genomes. If applicable, cross-referencing with the NCBI's UniGene database [18] was made possible *via* the Gene symbol attribute which can be found in many databases (Fig. 1). The final data set consisted of protein entries

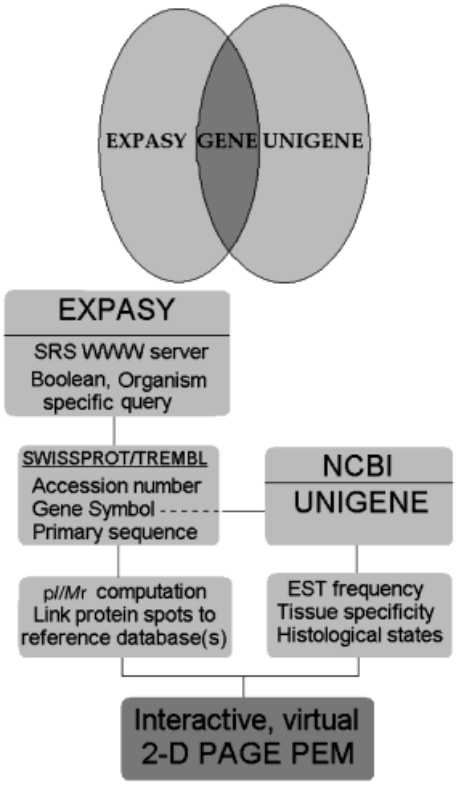


Figure 1. Data mining work flow. In principle, NCBI and SWISS-PROT entries share a common user-supplied Gene symbol field. Perl scripts were written to parse the organism/tissue output in both databases and scan them and organize it according to the common thread.

whose sequences have either actually been experimentally determined and submitted or predicted in reference to the ORF's from the corresponding genome.

3 Results

3.1 Analysis

Using the values in Table 1, and the primary sequences found in the latest release of SWISS-PROT/TrEMBL for humans, Perl scripts were written to parse the data into a tab-delimited format. The resulting plot of p*I versus* the molecular mass, yields a theoretical 2-D PAGE map with a striking bimodal distribution (Figs. 2 and 4). A total of 86 518 inferred or experimentally determined peptides were included in this calculation. One obvious feature of this map is the presence of a region seemingly devoid of proteins centered on pH 7.4–7.5. As shown in Fig. 3, this pattern is by no means unique to *Homo sapiens* and has been reported for other organisms [18–20]. However, to our knowledge, this distribution has never been reported with the currently available human genome and proteome

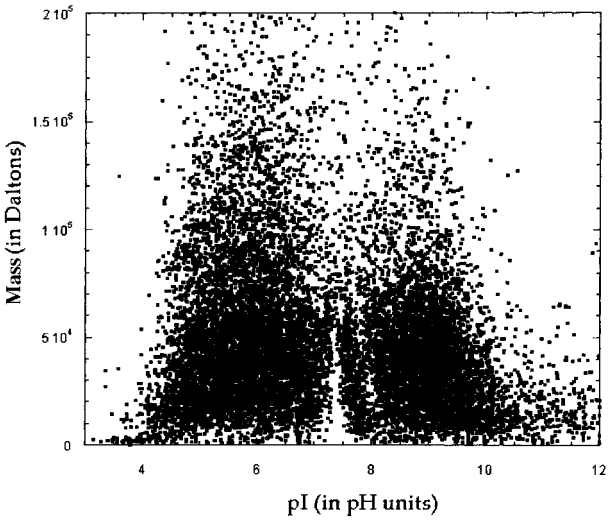


Figure 2. Proteins whose sequences were either experimentally determined or inferred are extracted for *Homo sapiens* from the latest combined SWISS-PROT-TrEMBL databases and plotted in the broadest range typically encountered in 2-D PAGE maps.

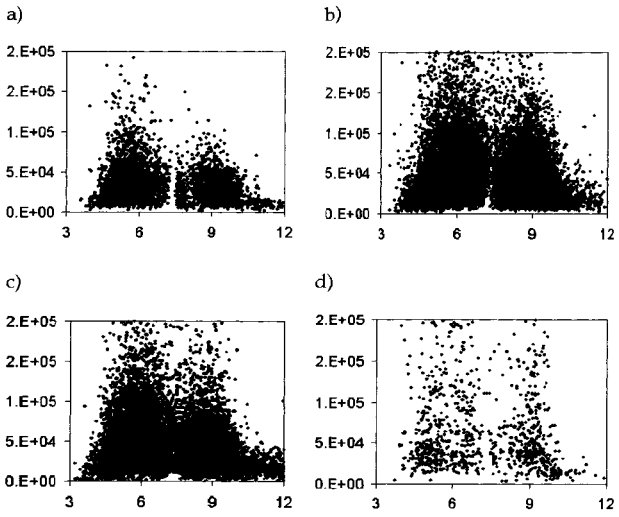


Figure 3. Other organisms surveyed from the same databases display a similar bimodal pattern a) *E-coli*; b) *C. elegans*; c) Mouse; d) *Plasmodium falciparum*.

data. The biochemical justification most often advanced in explanation of this observation is that the majority of proteins would tend to naturally precipitate out of solution around the cytoplasmic pH of approximately 7.2. The p*I* is the pH for which the protein charge is overall neutral. It therefore represents the point of minimum solubility due to the absence of electrostatic repulsion, resulting in maximum aggregation.

While this provides an explanation for experimental 2-D PAGE maps, we must remember that no such correction was incorporated in the modeling. What then is the basis for the separation of proteins into acidic and basic domains in computed p*I*/M_r charts? In our efforts to answer these questions, we carried out a simulation whereby groups of 1545 peptides varying in length from 50 to 600 amino acids (AA), in increments of 10 were randomly generated. This brings the total number of simulated sequences to 86 520 *versus* 86 518 real peptides extracted from current databases, thereby improving the prospects of any meaningful comparative statistics. As mentioned earlier, the calculation of the p*I* values is carried out iteratively. The pK of a peptide is calculated by tallying the contributions to the charge from the *N*-terminus, the *C*-terminus and the internal portion of the peptide. As can be observed in Fig. 4, the resulting simulated p*I*/M_r distribution is strikingly similar to that adopted by the extracted sequences. While this may seem surprising at first, given the total absence of bias in both the lengths and content of the peptides used for the simulation, it is in

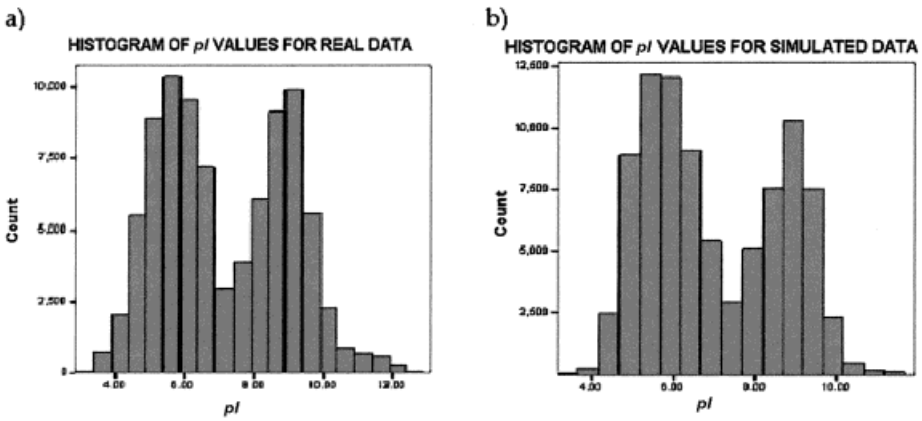


Figure 4. (a) Histogram of p*I* "real" values extracted from the latest combined release of SWISS-PROT/TrEMBL, for *H. sapiens*. The data points are grouped in bins that are 0.25 pH units wide. (b) Histogram of p*I* values, simulated data for each group, 1543 sequences are randomly simulated ranging from 50–600 amino acids in length, in increments of 10.

fact a direct consequence of the constraints imposed by a limited proteomic alphabet of twenty amino acids with distinct pK's roughly half of which are either acidic or basic (Table 1).

In fact, only seven internal amino acids make non-zero contributions to the p*I* of the peptide. These seven amino acids are: cysteine, aspartic acid, glutamic acid, histidine, lysine, arginine and tyrosine. It is reasonable to suspect that a high percentage of the variation in the calculated p*I* values of the simulated data would be modulated by the representation of these seven amino acids as the majority of the contribution to the charge comes from the internal portion of the peptide. To investigate the actual contribution of these seven amino acids in determining an overall p*I* value, a multiple regression model was developed using the adjusted numbers of these seven amino acids as predictor variables and the p*I* value as the dependent variable. The adjusted count for an amino acid is equal to the actual number of times the amino acid is found in the peptide divided by the length of the peptide. The adjusted counts will be denoted as follows: aR = adjusted count for arginine; aC = adjusted count for cysteine; aD = adjusted count for aspartic acid; aE = adjusted count for glutamic acid; aK = adjusted count for lysine; aH = adjusted count for histidine; aY = adjusted count for tyrosine.

The regression model in question uses the linear, quadratic and cubic powers for each adjusted number of the seven amino acids that contribute to the p*I* calculation when they are part of the interior of the protein. A total of 21 independent variables were employed in the regression analysis. This analysis yields a multiple correlation factor R of 0.931. The coefficient of determination (the square of the multiple R) gives the proportion of the total variance in the dependent variable accounted for by the set of independent variables in a multiple regression model. For the model in question, 0.866 is the square of the multiple R. Consequently, 86.6% of the total variation in the p*I* values was accounted for by the aforementioned seven amino acids. The simulation result confirms the hypothesis that the total number of these seven amino acids is the key factor in explaining the p*I* value of a peptide.

The predicted p*I* score in the regression model is denoted as p*I*' and it is the dependent (criterion) variable in the regression model. The equation for the regression model is:

$$pI' = a + \Sigma b_i X_i \quad (2)$$

where a is the intercept of the model, b_i is the partial slope for the ith predictor in the model and X_i is the ith predictor in the model. There will be 21 different predictors in the model: seven linear terms (aR, aC, aD, *etc.*), seven quadratic terms (aR^2, aC^2, aD^2, *etc.*) and seven cubic terms (aR^3, aC^3, aD^3, *etc.*). All parameters were estimated by ordinary least squares using the SPSS 8.0 computer package [21].

The coefficient of determination or R^2 for the model is the proportion of variance of the p*I* values accounted for by the regression model. It is equal to the sum-of-squares regression divided by the total sum-of-squares:

$$R^2 = \frac{\Sigma pI' - \text{Mean}(pI)^2}{\Sigma(pI - pI')^2} \quad (3)$$

In order to increase the analytical value of Virtual2D to the scientific community, interactivity is built into these plots by implementing the following features (displayed in Fig. 5): (i) accessibility from the WWW (http://proteom.

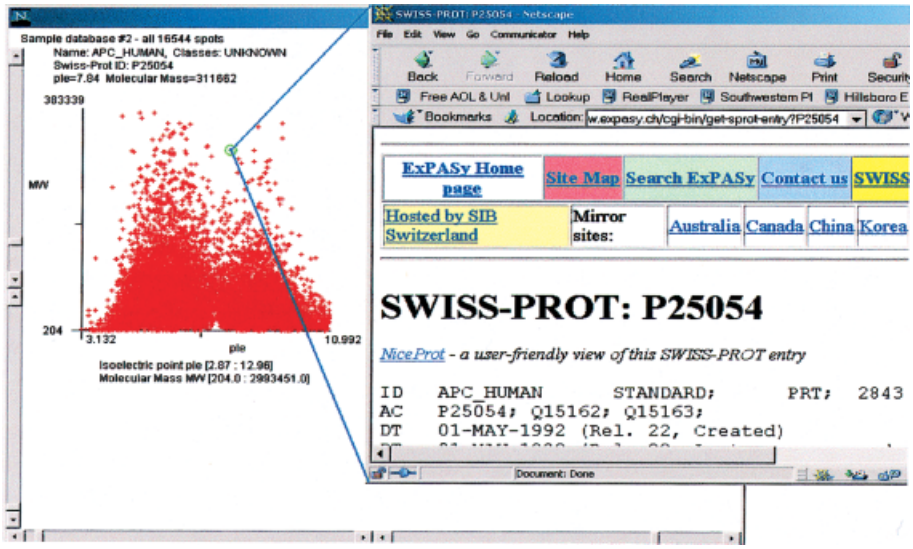


Figure 5. On the fly interaction and identification. By using the controls, one can zoom in on a particular area. Simply moving the mouse over or clicking on any spot will either display a short description or bring up comprehensive information from the hyper-linked web server of choice (Protplot uses Java code modified from MicroArray Explorer).

ncifcrf.gov/); (ii) zoom and click features; (iii) hyperlinks between each data point and popular databases (SWISS-PROT, NCBI, *etc.*). Depending on the actual application, these features were implemented in JAVA by modifying source code used in MAE (MicroArray Analysis Explorer) [22] or PtPlot [23], two versatile web-aware display programs.

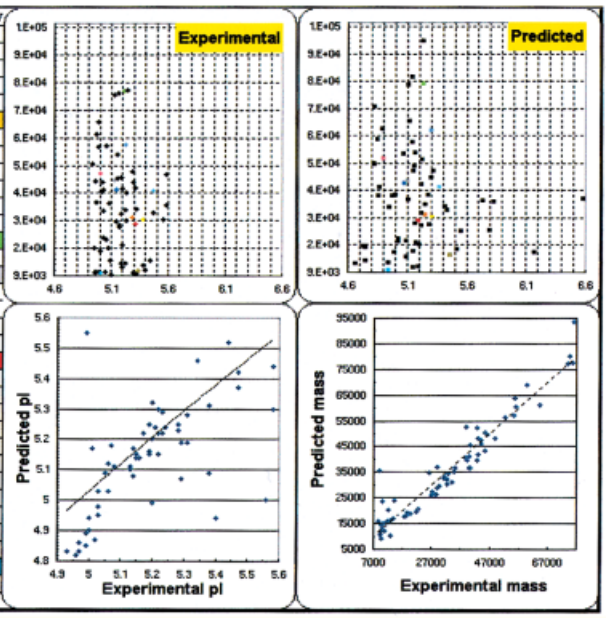
4 Discussion

4.1 Comparison with experimental data

We compared our computed p*I*/M_r values against those reported experimentally in two cases. In the first example, a high resolution map for *Escherichia coli* obtained over a narrow pH range (4.5–5.5) was used. Theoretical values of p*I*/M_r were computed and compared to the experimental observations. Landmarks provided by reference proteins whose characteristics were independently confirmed can be used to calibrate positions over the entire area of the image. p*I*, molecular masses and relative intensities can then be determined by interpolation for all detected protein spots (Fig. 6a). One is able to identify and select a reduced list of proteins whose predicted p*I*/M_r values are fairly close to their experimentally determined counterpart. This minimally distorted "constellation" set, displayed in Fig. 6b can then be used in principle to "warp" the two gels, thereby aligning the experimental map on top of the theoretical one.

a

Exp pI	Exp Mw	Ac	Pred pI	Pred Mw
4.96	9467	P00274	4.67	11675
5.44	13981	P00501	5.52	16971
5.03	39792	P00574	4.98	36512
5	45494	P00824	4.9	50194
5.34	10379	P00832	5.46	14937
5.38	28867	P00928	5.31	28724
4.99	64444	P02349	4.89	61158
5.1	8830	P02418	6.17	15769
5.31	40279	P02917	5.28	36744
5.19	31192	P02927	5.25	33368
5.23	38521	P02928	5.22	40707
5.58	45310	P02990	5.3	43182
5.21	75514	P02996	5.24	77450
5.22	32484	P02997	5.22	30292
5.13	43069	P04384	5.1	41952
4.97	59766	P04475	4.83	68984
5.56	28995	P04816	5	36911
5.13	73960	P05055	5.11	77101
4.99	9991	P05082	5.55	23586
5.31	27332	P05194	5.19	27467
5.19	44130	P05313	5.16	47522
5.22	12998	P05380	5.15	10387
4.99	55284	P06139	4.85	57138
5.32	9085	P06977	6.58	35401
5.22	56152	P08142	5.3	60441
5.17	26187	P08186	5.74	34916
5.15	43951	P08200	5.15	45757
5.2	40116	P08324	5.32	45524
4.95	18479	P08837	4.73	18120
5.17	26881	P09551	5.22	25785
5.32	32815	P11096	5.83	34358
5.14	39792	P11665	5.08	40987
5.16	74539	P13029	5.14	80024
5.06	55572	P16659	5.12	63734



b

Pi(Pred)	M(Pred)	Pi(exp)	M(Exp)	Ac
4.9	50194.23	4.99892	45494	P00824
5.31	28724.16	5.38	28936	P00928
4.89	61158.07	4.99	64780	P02349
5.24	77450.11	5.20869	76087	P02996
5.1	41951.63	5.13469	43079	P04384
5.3	60440.57	5.22236	56329	P08142
5.15	45756.71	5.15158	43958	P08200
5.22	25784.99	5.17088	26913	P09551
5.08	40987.03	5.1371	39816	P11665
5.42	32664.16	5.46889	31435	P16703
5.03	19572.36	5.06431	21653	P17288
4.95	32286.01	5.03085	32271	P25715
5.03	20630.25	5.02667	22398	P26427
4.94	9065.24	4.99784	9673	P26605
5.11	35088.05	5.08483	34775	P30148
4.75	17704.12	5	17510	P37901
5.17	26232.61	5.13871	28451	P39182

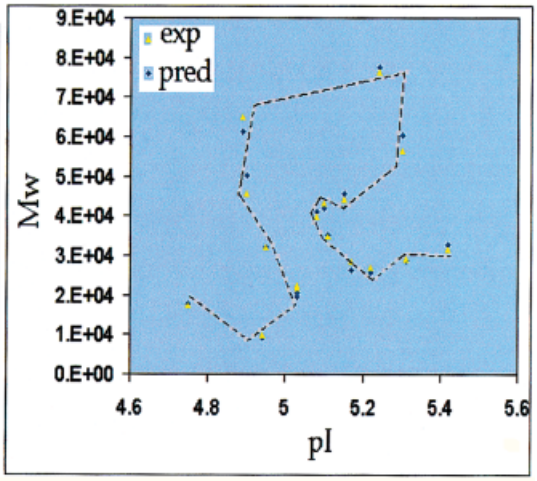


Figure 6. (a) Comparison of the values of IEF points and molecular mass extracted from a high resolution *E. coli* 2-D PAGE map downloaded from SWISS-2-D PAGE and those computed in this work. In the two upper charts, a small number of corresponding data points from each set have the same color for a quicker visual inspection. (b) For a small subset of proteins, computed p*I*/M_r values are fairly close to the experimental counterparts, providing a "constellation" of reference points that can be used for warping.

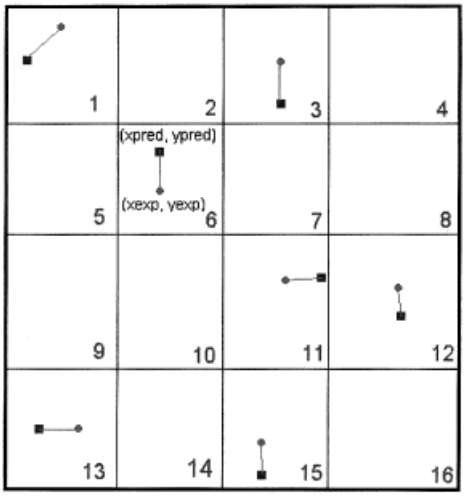


Figure 7. The warping of a 2-D PAGE map on a computed p*I*/M_r chart can be achieved by dividing it in areas surrounding each pair of experimental (●) and predicted (■) landmarks and applying to all the protein spots belonging in a particular neighborhood the necessary local translation to transform the coordinates (X_{pred}, Y_{pred}) to (X_{exp}, Y_{exp}).

As an example, one can imagine dividing up the gel into several regions around each one of these pairs of spots so that for any given region, the local experimental landmark (closed circle) will be transformed to its predicted counterpart (closed square) by a translation specific to that neighborhood (Fig. 7). Any experimental spot (including the landmark) within region 1 for instance will undergo the same local translation defined by:

$$X_{pred} = X_{exp} + \triangle X_1 \quad (4)$$
$$Y_{pred} = Y_{exp} + \triangle Y_1$$

where $\triangle X1$ and $\triangle Y1$ are the components of the local translation needed to bring an experimental landmark onto its predicted counterpart. If the spot happens to be in region 3, then

$$X_{pred} = X_{exp} + \triangle X_3 \quad (5)$$
$$Y_{pred} = Y_{exp} + \triangle Y_3$$

and so on.

For those areas without designated landmark such as region 2, one can interpolate using the translations from the surrounding neighborhoods.

$$X_{pred} = X_{exp} + \triangle X_2 \text{ where } \triangle X_2 = \quad (6)$$
$$= (\triangle X_1 + \triangle X_3 + \triangle X_6)/3$$
$$X_{pred} = Y_{exp} + \triangle Y_2 \text{ and } \triangle Y_2 =$$
$$= (\triangle Y_1 + \triangle Y_3 + \triangle Y_6)/3$$

The outcome of this 2-D alignment is not a trivial task as it is a function of several factors including the resolution of the experimental gel (the higher, the better) as well as the number and spatial distribution of landmark reference points. It involves working out the transformations that reflect the local distortions of the gel. Several software packages [24–26] currently existing on the market offer robust and flexible spot detection from many popular image file formats coupled with sophisticated statistical and warping tools.

In the second example, we (arbitrarily) selected and downloaded from SWISS-2D PAGE a map of human colorectal epithelia cells [27]. Figure 8 depicts the overlap

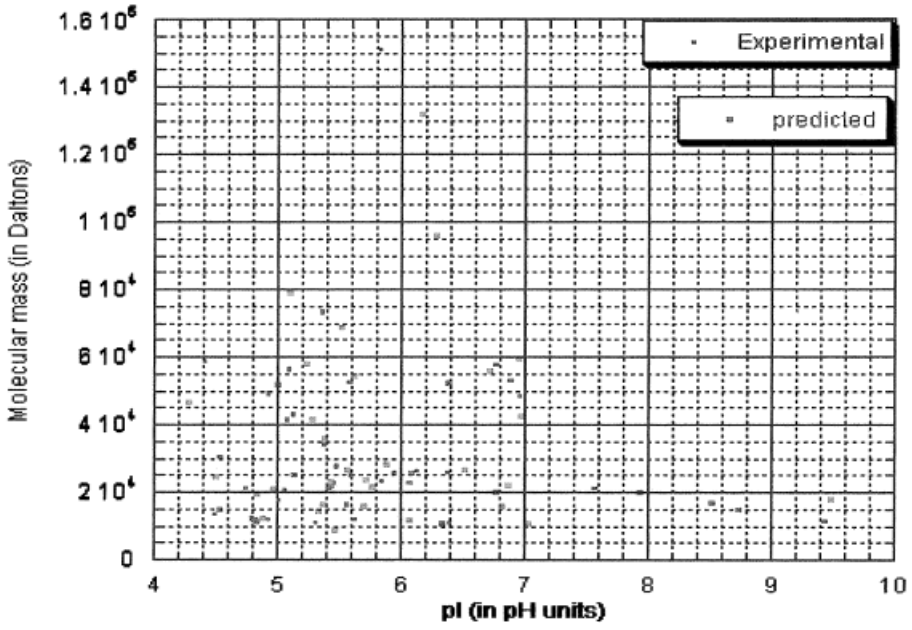


Figure 8. Overlap of p*I*/M_r experimental (●) and theoretical values (■) for spots identified in a 2-D PAGE map of human colorectal epithelial [27] obtained from SWISS-2D PAGE.

of observed and corresponding computed p*I*/M_r values for 40 proteins. A quantitative measure of the discrepancy between the two data sets can be obtained by using the relative shift (r.s) of a protein spot between experimental and theoretical values:

$$r.s = [(\triangle pI/pI_{exp})^2 + (\triangle M_r/M_{r\,exp})^2]^{1/2}$$

where

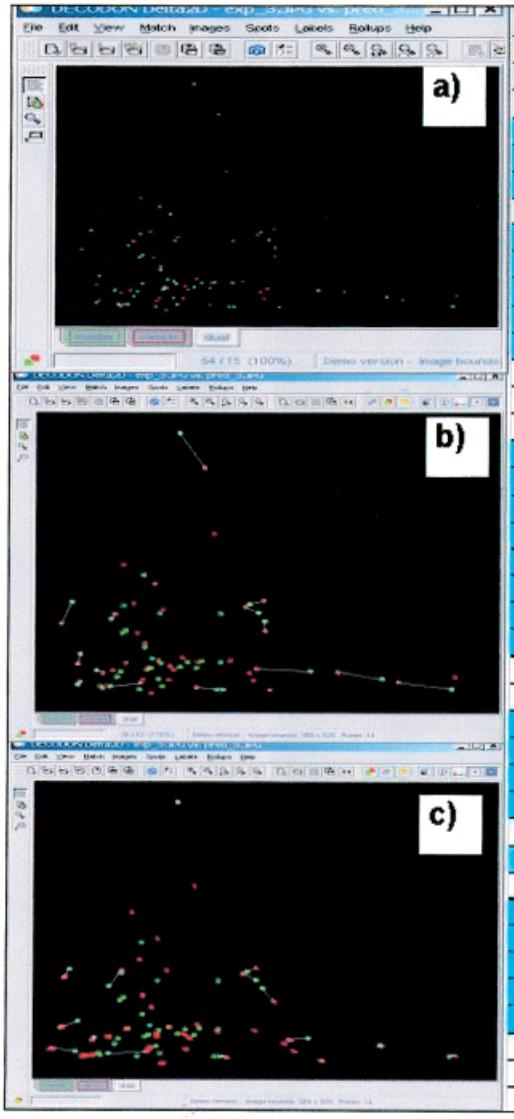
$$\triangle pI = pI_{exp} - pI_{pred} \text{ and } \triangle M_r = M_{r\,exp} - M_{r\,pred} \quad (7)$$

Despite the broad nominal intervals for p*I* (4–8 pH units) and M_r (0–200 kDa), more than 66% of the predicted values have a relative shift less than or equal to 0.12 compared to their observed counterpart.

However, one must still face the reality of the numerous modification types occurring co- and post-translationally which can severely alter the electrophoretic mobility of the proteins affected. As can be seen in Fig. 9, while relatively small local differences can be easily be reconciled, no amount of warping will be able to totally and correctly align a collection of computed p*I*/M_r data points onto a set of experimentally determined protein spots without individually identifying and incorporating the aforementioned corrections in the computation of these attributes.

4.2 Correlation to expressed sequence tag data

The relationship between mRNA levels and protein expression has been explored by a number of investigators [28]. The data type generated by VIRTUAL2D, together



Exp pI	Exp Mw	Pred pI	Pred Mw	r.s
4.41	58700	4.29	46466	0.210
4.5	13360	4.53	15020	0.124
4.54	30228	4.5	24633	0.185
4.75	21013	4.84	19595	0.070
4.8	12166	4.88	12513	0.033
4.84	11163	4.82	11606	0.040
4.93	12076	5.34	14585	0.224
4.93	49118	5	51769	0.056
5.01	17462	5.37	16394	0.094
5.08	20750	4.97	21014	0.022
5.08	41448	5.29	41605	0.042
5.1	56317	5.24	57963	0.040
5.13	42930	5.1	79227	0.846
5.31	11036	9.48	17819	0.397
5.31	11036	5.46	8960	0.190
5.36	73173	5.51	68858	0.065
5.38	34062	5.38	35890	0.054
5.42	21346	5.44	22222	0.041
5.46	22765	5.44	23225	0.021
5.48	27623	5.56	28629	0.039
5.56	16289	5.7	15805	0.039
5.58	52522	5.61	54265	0.033
5.8	24976	5.13	24976	0.084
5.62	12121	7.03	10612	0.280
5.83	150908	6.17	132013	0.138
5.85	23060	5.71	23760	0.033
5.95	25834	5.88	26355	0.104
6.07	22824	5.77	21468	0.077
6.13	26432	6.08	25853	0.023
6.34	10683	6.81	15867	0.491
6.38	25826	6.51	26538	0.034
6.39	52160	6.28	95798	0.837
6.4	11038	6.07	11731	0.081
6.77	57695	6.35	59625	0.043
6.89	52776	6.71	56006	0.067
6.96	48310	6.97	42769	0.115
7.57	21182	6.86	22204	0.105
8.52	16844	7.93	19974	0.198
8.52	16844	6.76	20159	0.265
9.44	11331	8.73	15126	0.343

Figure 9. (a) Overlap of spots identified in 2-D PAGE map of human colorectal epithelial cell line (green) and theoretically computed (red). (b) Several pairs of corresponding experimentally predicted spots are connected to reflect the translations. (c) A global warping attempts to bring the computed value closer to the corresponding observed member of the pair. While in some cases, an almost exact local alignment is achieved, in many instances the differences caused by post-translation modifications are simply too large to successfully align them. This analysis was carried out using a demonstration version of the Delta-2D package [26].

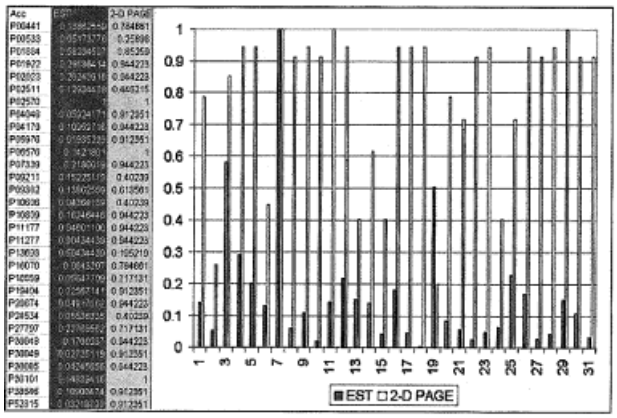


Figure 10. Bar graph comparison of relative frequency of EST for the whole *H. sapiens* organism *versus* protein expression levels derived from integrated optical densities of corresponding spots in 2-D PAGE map of colorectal epithelial cells.

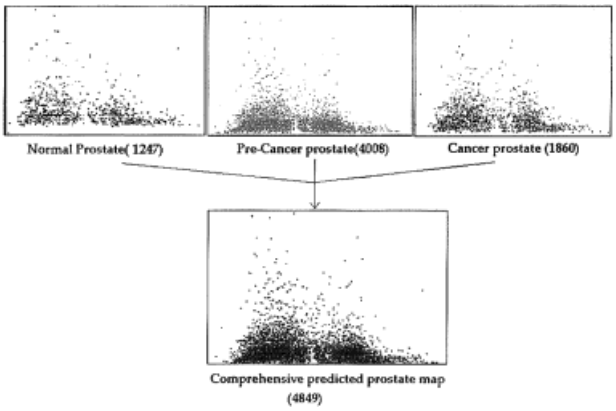


Figure 11. Predicted p*I*/M_r maps of proteins expressed in the prostate gland. Information for each histological state is derived from EST data in the corresponding CGAP Library (281-Normal, 282-PreCancer, 283-Normal). A comprehensive prostate specific p*I*/M_r map would correspond to the union of all three histological states.

with sequence databases such as UNIGENE [29, 30], provides an excellent opportunity to directly address this question. Analysing the high resolution 2-D PAGE map of colorectal epithelium cells mentioned in Section 4.1, the optical density and area of each of these spots were used to determine the quantity of each protein relative to that of the whole image. In the event that several isoforms of the same protein are present in the gel (due to modifications), their contributions were summed. At the same time, the corresponding number of EST 'hits' for each protein was determined from a comprehensive survey of human mRNA libraries.

The results shown in Fig. 10 indicate an overall preserved trend (*i.e.*, when the relative abundance of an mRNA increases, so does the expression level of the corresponding protein as seen on the 2-D PAGE map, albeit not always linearly). The lack of a more stringent correlative relationship between mRNA levels and protein abundance is most likely due to the influence of post-translational gene regulation (*i.e.*, mRNA stability, translational efficiency and protein stability).

4.3 The hunt for disease markers, application to cancer

One approach to uncover biological markers associated with the early onset of disease is to detect statistically significant differences in protein expression levels and establish correlations with disease outcome or onset. Prostate cancer is diagnosed every 2 ½ min with approximately 200 000 new cases diagnosed each year in North America [31]. It is the most common cancer in the United States among men. The Cancer Genome Anatomy Project (CGAP) [32, 33] is an interdisciplinary program established and administered by the US National Cancer Institute (NCI) to generate the information and technological tools needed to decipher the molecular anatomy of the cancer cell.

The advent of laser capture microdissection (LCM) [34, 35] has provided the ability to procure a pure population of cells that would give the most accurate results of expression profiling as a function of tissue state. We therefore decided to focus on LCM-generated, non-normalized CGAP libraries of EST from the UniGene database to minimize any bias inherent to the cloning method used.

Of the libraries satisfying these criteria, three prostate-related sets (281, 282 and 283) provide substantially more data points and therefore a significantly better suited ensemble upon which to draw any meaningful statistical conclusions. In addition, these samples cover all three histological states of the disease and originate from epithelial cells taken from the periurethral zone of the prostate gland of the same patient.

By compiling the information from those ESTs that have already been assigned to a specific gene, we can then deduce the associated protein(s) and their respective properties leading to a predicted p*I*/M_r chart for each library. A comprehensive prostate map would then consist of the union of all three maps. The results are shown in Fig. 11. By combining the data obtained from similar libraries (same tissue and histological state) and further tracking the relative frequencies of occurrence of these ESTs and assigning proportional intensities, one is able

to build gray-scale theoretical protein expression maps that are both tissue and histology specific. Figure 12 depicts the profiles of all gene products expressed in the prostate gland in various cancer states. (The results of this exhaustive survey will be thoroughly described in a separate paper.)

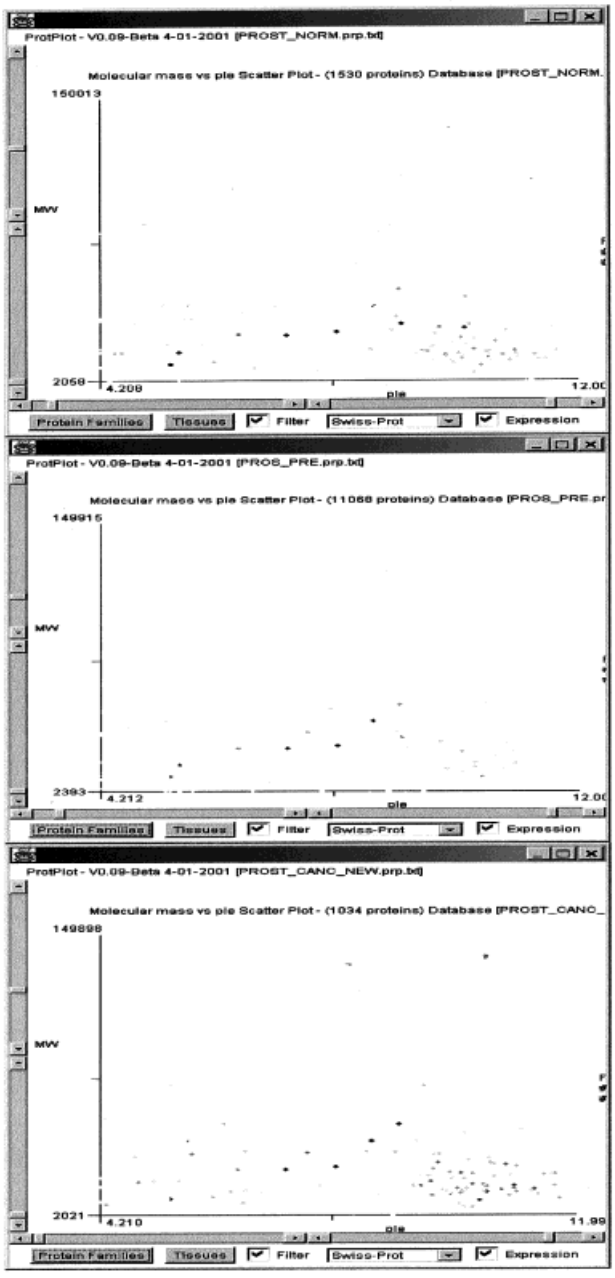


Figure 12. Virtual 2-D expression maps of prostate derived from pooling 13 UNIGENE libraries that span three histological states: Normal (top), Pre-Cancer (middle) and Cancer (bottom). User-controlled handles for zooming, web-server and tissue selection. The gray-scale and relative spot intensities reflect the corresponding EST frequencies in the pooled data set.

5 Concluding remarks

We propose in this work a simple, yet plausible, explanation for the bimodal p*I*/M_r distribution observed experimentally and predicted theoretically. To date, theoretical maps for ninety-two organisms/proteomes compiled have been computed and deposited within this database (accessible on the web at http://proteom.ncifcrf.gov/). It is designed to be a reference tool to assist investigators in the putative assignment of proteins in whole genome complements. It offers the ability to optimize a narrow pH range, prior to actually running a 2-D PAGE experiment, according to the expected attributes (p*I*, M_r) of the proteins of interest. In addition, one is able to predict the expression and approximate location of the unmodified isoform of gene products. If their expression level is otherwise too low to be detected by traditional Coomassie or silver staining, one may be able, through enrichment techniques and enhanced imaging and data integration, to improve the sensitivity of 2-D PAGE. This is of particular importance in identifying and quantitatively profiling biomarkers associated with the early onset of diseases such as cancer through the various histological states and is the focus of ongoing efforts. Eventually, the workflow on the right-hand side of Fig. 13 will enable the user to input a file representing an entire proteome and ultimately a full-blown genome and produce the desired map.

Cross-referencing between the various databases was no small task even though in principle, entries in ExPASy's SWISS-PROT and NCBI's UniGene share a common key (Gene symbol). In our experience, more than a third of the assigned names were different enough to frustrate any

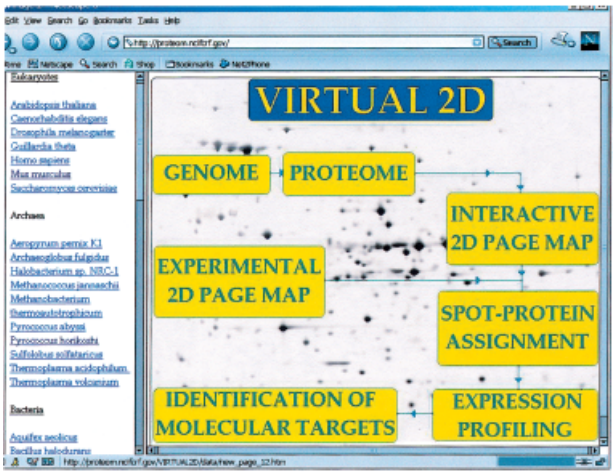


Figure 13. A snapshot of the screen display of VIRTUAL2D accessible at http://proteom.ncifcrf.gov. Protein expression maps computed for fifty-three organisms/proteomes using data obtained from the European Bioinformatics Institute [36] can be displayed by clicking on any of the entries on the left.

attempts to automate the harvesting of overlapping information. This highlights the need for better quality control in these repositories and the enforcement of at least one key/thread common to all existing life-science databases.

The authors wish to gratefully acknowledge Christine Hoogland (Swiss Institute of Bioinformatics) and Chris Santos (CIT, NIH) for assistance with algorithms, Dr. Lucas Wagner (NCBI) for concrete suggestions, Drs. Joseph Kates (SAIC-Frederick) and Patrick O'Farrell (UCSF) for useful discussions. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organization imply endorsement by the US Government. This project has been funded with Federal funds from the National Cancer Institute, National Institutes of Health, under contract No NO1-CO-12400.

Received March 20, 2002

6 References

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

[2] O'Farrell, P. Z., Goodman, H. M., O'Farrell, P. H., *Cell* 1977, *12*, 1133–1141.

[3] Aebersold, R., Rist, B., Gygi, S. P., *Ann. NY Acad. Sci.* 2000, *919*, 33–47.

[4] Bussow, K., *Trends Biotechnol.* 2001, *19*, 328–329.

[5] Fivaz, M., Vilbois, F., Pasquali, C., van der Goot, F. G., *Electrophoresis* 2000, *21*, 3351–3356.

[6] Kriegel, K., Seefeldt, I., Hoffmann, F., Schultz, C. *et al., Electrophoresis* 2000, *13*, 2637–2640.

[7] Dihazi, H., Kessler, R., Eschrich, K., *Anal. Biochem.* 2001, *299*, 260–263.

[8] Angelis, F. D., Tullio, A. D., Spano, L., Tucci, A. J., *Mass Spectrom.* 2001, *36*, 1241–1248.

[9] Weiller, G. F., Djordjevic, M. J., Caraux, G., Chen, H., Weinman, J. J., *Proteomics* 2001, *12*, 1489–1494.

[10] Wulfkuhle, J. D., McLean, K. C., Paweletz, C. P., Sgroi, D. C. *et al., Proteomics* 2000, *10*, 1205–1215.

[11] WORLD-2DPAGE: http://www.expasy.ch/ch2d/2d-index.htm

[12] http://www.expasy.ch/cgibin/map2/def?HEPG2_HUMAN

[13] Görg, A., Obermaier, C., Boguth, G., Harder, A. *et al., Electrophoresis* 2000, *6*, 1037–1053.

[14] Bjellqvist, B., Sanchez, J. C., Pasquali, C., Ravier, F. *et al., Electrophoresis* 1993, *14*, 1375–1378.

[15] http://www.lionbioscience.com/solutions/srs

[16] Bairoch, A., Apweiler, R., *Nucleic Acids Res.* 2000, *28*, 45–48.

[17] pI/Mw is part of ExPASY's proteomics tools: http://www.expasy.ch/tools/pi_tool.html

[18] http://www.ncbi.nlm.nih.gov/UniGene

[19] VanBogelen, A. R., Abshire, Z. A., Moldover, B., Olson, R. E., Neidhardt, C. F., *Electrophoresis* 1997, *18*, 1243–1251.

[20] Bairoch, A., in: Wilkins, M. R., Williams, K. L., Appel, R. D., Hochstrasser, D. F. (Eds.), *Proteome Research: New Frontiers in Functional Genomics*, Springer-Verlag Berlin Heidelberg, 1997, pp. 93–148.

[21] http://www.spss.com/

[22] Lemkin, P. F., Thornwall, G., Walton, K., Hennighausen, L., *Nucleic Acids Res.* 2000, *22*, 4452–4459; http://www.lecb.ncifcrf.gov/MAExplorer/

[23] http://ptolemy.eecs.berkeley.edu/java/ptplot/

[24] http://www.www.expasy.ch/melanie/

[25] http://www.2dgels.com/

[26] http://www.decodon.com/

[27] Reymond, M. A., Sanchez, J.-C., Hughes, G. J., Riese, J. *et al., Electrophoresis* 1997, *18*, 2842–2848.

[28] Anderson, L., Seilhamer, J., *Electrophoresis* 1997, *18*, 533–537.

[29] Schuler, G. D. *J. Mol. Med.* 1997, *75*, 694–698.

[30] Boguski, M. S., Schuler, G. D., *Nat. Genet.* 1995, *10*, 369–371.

[31] Information about the origin, statistics and current research methodologies in most commonly diagnosed types of cancers can be found at http://nci.nih.gov

[32] Strausberg, R. L., Dahl, C. A., Klausner, R. D., *Nat. Genet.* 1997, *15*, Spec, 415–416.

[33] http://cgap.nci.nih.gov/

[34] Krizman, D. B., Chuaqui, R. F., Meltzer, P. S., Trent, J. M. *et al., Cancer Res.* 1996, *56*, 5380–5383.

[35] Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F. *et al., Science* 1996, *8*, *274*, 998–1000.

[36] http://www.ebi.ac.uk/