Au: please mention all figures in text

6 Proteome Knowledge Bases in the Context of Cancer

Figures menetioned out of order

Djamel Medjahed and Peter A. Lemkin

CONTENTS

1-57444-480-8/05/$0.00+$1.50

6.1 INTRODUCTION

The origin of most cancers can be often traced to a single transformed cell.[1] The evolution of the disease follows a yet-to-be completely understood pathway of molecular transformations occurring at both genomics and proteomics levels as depicted in Figure 6.1. Most cancers show a significant preponderance to statistically originate from well-defined part of their respective organs. It is then only normal that investigations to identify biomarkers indicative of the early onset of the disease be focused on these organ-specific regions.

This point was elegantly demonstrated by Page et al.[2] in a careful experiment, where they used magneto-immuno-chemical purification methods to extract pure cell populations and compare the protein expression observed in experimental two-dimensional poly-acrylimide gel electrophoresis (2D PAGE) maps obtained from normal, milk-producing luminal epithelial cells exhibiting a tendency to exhibit carcinomas vs. outer, myoepithelial cells as described in Figure 6.2.

This thorough characterization was achieved by using a combination of enabling technological platforms, some of which are listed in Figure 6.3, which allowed them able to flag a number of proteins exhibiting a significant differential expression between the two types of cells and therefore warranting a closer evaluation of their

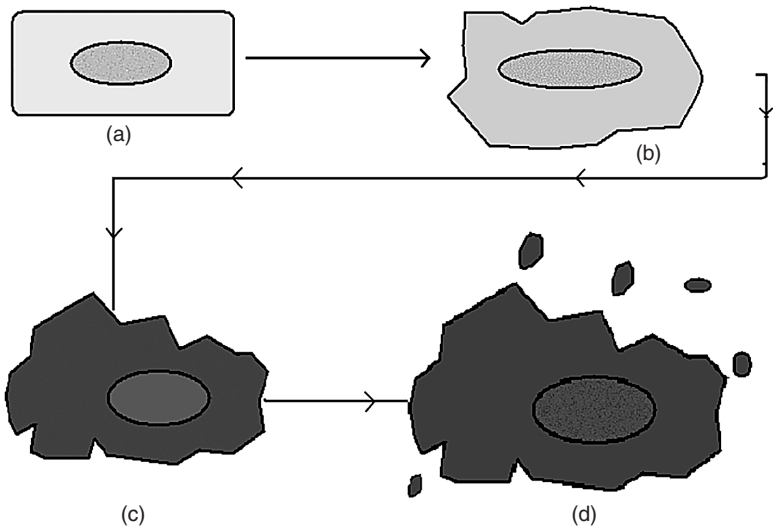


FIGURE 6.1 Illustration of the progressive evolution from a normal cell (a) to precancer (b and c), and finally the cancerous state (d).

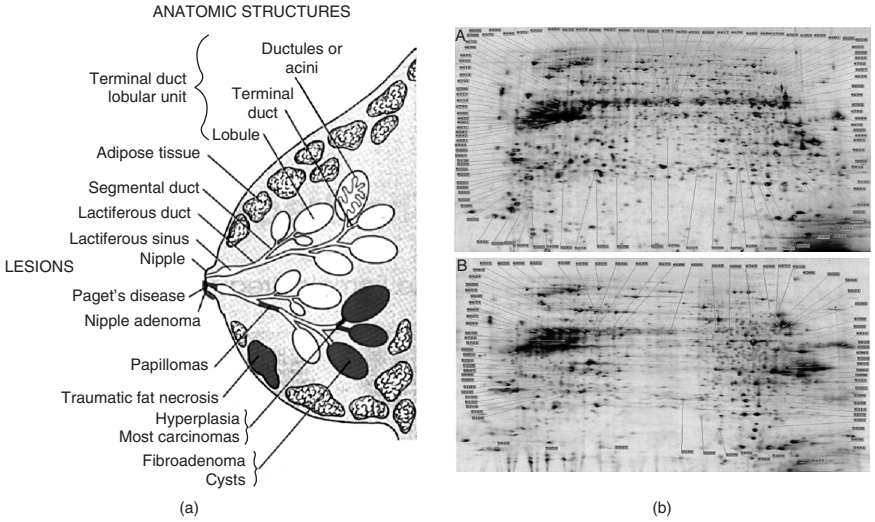


FIGURE 6.2 Dissection of the organ breast. (a) The lobular-alveolar regions colored in pink show the nature of lesions statistically originating from them. (b) Comparison of protein expression profiles in the inner, epithelial luminal (A) that account for 95% of breast carcinomas vs. (B) outer, myoepithelial of healthy patients. The annotations are those of 51 proteins, which display more than twofold expression change between the two samples.

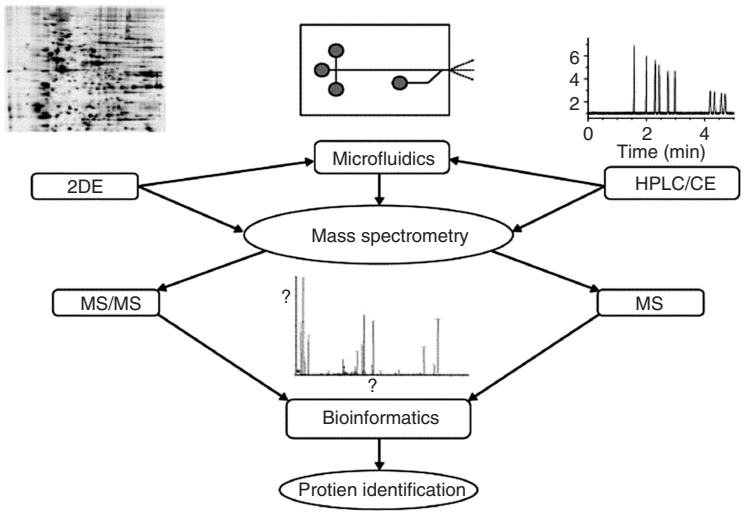


FIGURE 6.3 Technological platforms used in proteomic characterization.

potential as biomarkers of breast cancer. The time and costs involved in using these techniques can be quite prohibitive, particularly on a large scale.

This led to the initial motivation to address the need that either on a routine basis, or to establish optimal experimental conditions before hand, one might be interested in predicting the gene products likely to be detected in narrow ranges of isoelectric focusing point (pI) and molecular weight (Mw).

We believe that the initial search for cancer biomarkers can greatly benefit by formulating hypotheses developed from knowledge-based bioinformatic tools. This chapter will describe in some detail two such predictive databases whose development was at least in part motivated by these pressing issues.

6.2 VIRTUAL 2-D: A WEB-ACCESSIBLE PREDICTIVE DATABASE FOR PROTEOMICS ANALYSIS

Over the past three decades and thanks to continuous developments in chemistry,[3] automation, and data collection,[4–6] 2D PAGE[7,8] has evolved from a labor intensive, multiprocess protein separation method to becoming an integral part of most comprehensive proteomics efforts.[9–12] In particular, the advent of immobilized pH gradients[13] in the first dimension has ushered in an era where reproducible, high-resolution iso-electric focusing measurements can routinely be carried out, making it conceivable to predict from the primary sequence the equilibrating positions of proteins within a pH gradient. When solubilized with high concentrations of urea (8.5–10 M), proteins unfold and only the ionizable groups or those amino acids located at the N- or C-terminal 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 used to determine the isoelectric focusing point and molecular mass of a peptide can then simply be summed up as follows:

1. Scan the primary sequence of the peptide
2. Assign the pK of each contributing amino acid according to Table 6.1
3. Sum up all the mass contributions

The resulting Pi/Mw for the peptide is then given by the ratio of:

$$\{pK_{Cterm} + \Sigma_{int}\, pK_{int} + pK_{Nterm}\}$$

$$Pk_{tot} = (n - 2)$$

and

$$M_{rtot} = \Sigma_{\,i}\, M_t i \tag{6.1}$$

where the pI summation runs over all *n* contributing, internal amino acids.

Au: please supply table titles for all tables

TABLE 6.1

Au: Please reference all tables in text

Ionizable Group	PK	Molecular Mass
C-terminal	3.55	
N-terminal		
Met	7.00	132.994
Thr	6.82	102.907
Ser	6.93	88.88
Ala	7.59	72.88
Val	7.44	100.934
Glu	7.70	130.917
Pro	8.36	98.918
Internal		
Asp	4.05	116.89
Glu	4.45	130.917
His	5.98	138.943
Cys	9	104.94
Tyr	10	164.978
Lys	10	114.961
Arg	12	157.989
C-terminal side chain groups		
Asp	4.55	116.89
Glu	4.75	130.917

6.2.1 Database Mining

~~*Homo sapiens* were the first organisms that were~~ examined. The resulting plot of pI versus the molecular mass yields a theoretical 2D PAGE map with a striking bimodal distribution (~~Figure~~ ~~6.4a~~). 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 to 7.5.

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 pI 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 2D 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 pI/MW 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 AA, in increments of 10, were randomly generated. This brings the total number of simulated sequences to 86,520 vs. 86,518 real peptides

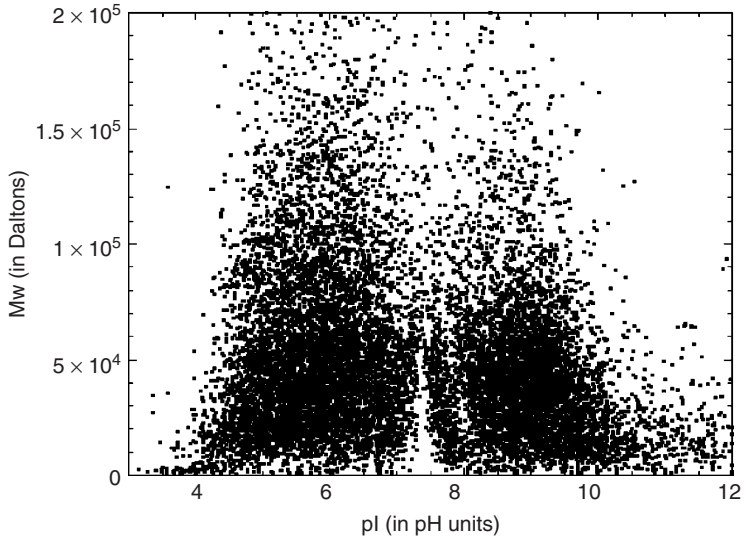
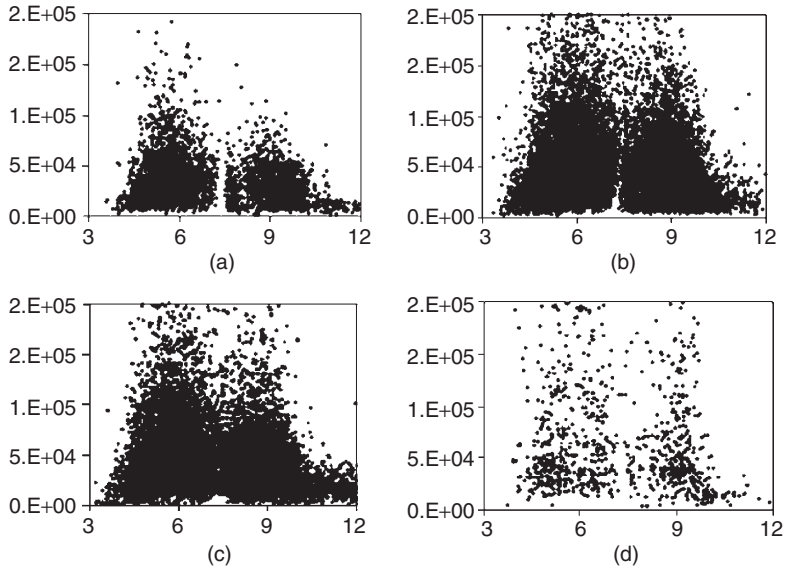


FIGURE 6.4 pI/Mw Map for *Homo sapiens*. To keep in line with the experimental limits encountered in practice, the pI/MW plot has been confined to less than 2×10^5 kD for the molecular mass and $3.0 < pI < 12.0$ for the isoelectric focusing point. As shown in Figure 6.5, this pattern is by no means unique to *Homo sapiens* and has been reported for other organisms.[21–23]

extracted from current databases, thereby improving the prospects of any meaningful comparative statistics. As mentioned earlier, the calculation of the pI 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 Figure 6.4, the resulting simulated pI/MW 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 fact a direct consequence of the constraints imposed by a limited proteomic alphabet of twenty amino acids with distinct pKs, roughly half of which are either acidic or basic (Table 6.1).

In fact, as is reflected in Table 6.1, only seven internal amino acids make non-zero contributions to the pI 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 pI 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 pI value, a multiple regression model was developed using the adjusted numbers of these seven amino acids as predictor variables and the pI value as the dependent variable. The adjusted count for an amino acid is equal to



Au: a and b meant?

FIGURE 6.5 pI/Mw Charts for (a and b) *E. coli*, c) mouse, and d) *Plasmodium falciparum*.

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 pI 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 .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, .866 is the square of the multiple R. Consequently, 86.6% of the total variation in the pI 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 is explaining the pI value of a peptide.

The predicted pI score in the regression model is denoted as pI′ and it is the dependent (criterion) variable in the regression model. The equation for the regression model is:

$$pI' = a + \Sigma b_i X_i \tag{6.2}$$

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: 7 linear terms (aR, aC, aD, etc.), 7 quadratic terms (aR^2, aC^2, aD^2, etc.), and 7 cubic terms (aR^3, aC^3, aD^3, etc.). All parameters were estimated by ordinary least squares using the SPSS 8.0 computer package.[15]

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

$$\Sigma pI' - <(pI)>)^2$$

$$R^2 = \Sigma(pI - pI')^2 \tag{6.3}$$

Au: lowercase "i" ok?

where $<(pI)> = \Sigma\, pi/N$

Unpredictable bottlenecks associated with Internet traffic and limitations in the size of the files that could be downloaded at any given time from the pI/Mw server force one to typically fragment the proteome of an organism into several smaller files no bigger than 2000 gene product entries. A Perl script was written to address this issue, and, when applied to organism-specific, curated proteome datasets in FASTA format downloaded from the European Bioinformatics Institute's web site, will output tab-delimited files of the molecular mass, pI, Swiss-Prot accession number and identification for each protein entry. 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 Figure 6.7).

Possibility to use the database on any JAVA-enabled computer
Pan, zoom, and click features
With Internet connection, hyperlinks between each data point and popular databases (Swiss-Prot, NCBI, etc.)

6.2.2 Comparison with Experimental Data

Computed pI/MW values were compared against those reported experimentally in two cases. In the first example, a high-resolution map for *E. coli* obtained over a narrow pH range (4.5–5.5) was used. Landmarks provided by reference proteins whose characteristics were independently confirmed can be used to calibrate positions over the entire area of the image. pI, molecular masses, and relative intensities can then be determined by interpolation for all detected protein spots (Figure 6.6a). A minimally distorted "constellation" consisting of proteins whose predicted pI/Mw values are fairly close to their experimentally determined counterpart, displayed in

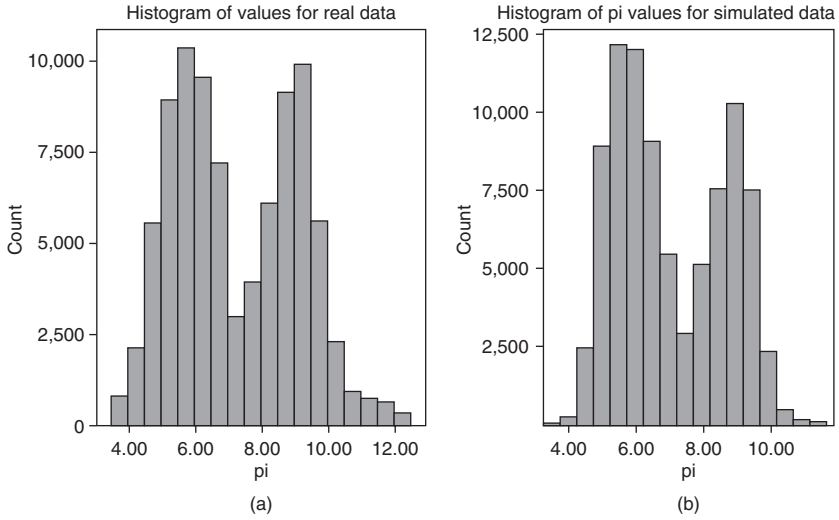


FIGURE 6.6 Side-by side comparison of ''pI/Mw'' histograms for *Homo sapiens*. (a) Computed using amino acid sequences from TrEMBL/Swiss-Prot vs. (b) randomly generated as described in the text.

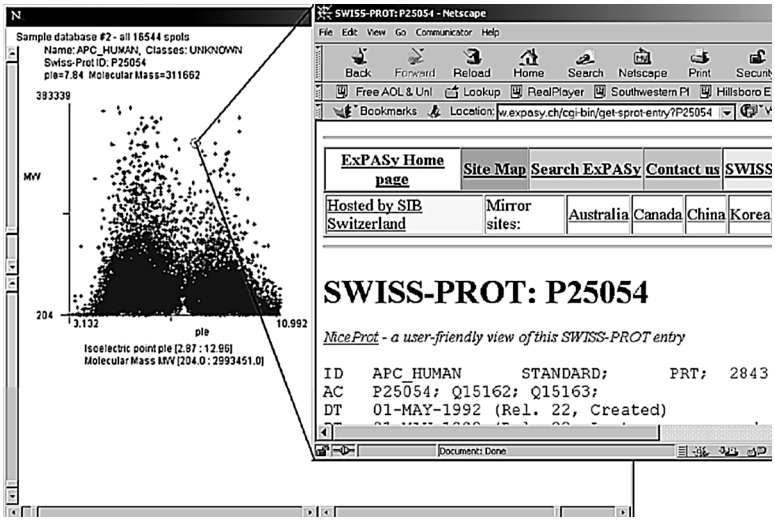


FIGURE 6.7 On-the- fly interaction and identification. By using the controls, one can zoom in on a particular area. Simply moving the mouse over or clicking on any spot will either display a short description or bring up comprehensive information from the hyperlinked web server of choice (Protplot uses Java code modified from MicroArray Explorer).

Figure 6.6b can then be used in principle to "warp" (align) the experimental gel onto the theoretical one.

To understand warping in its simplest form, 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 (brown circle) will be transformed to its predicted counterpart (blue square) by a translation specific to that neighborhood (Figure 6.7). Any experimental spot (including the landmark) within region 1, for instance, will undergo the same local translation defined by:

$$X_{\text{pred}} = X_{\text{exp}} + \Delta X_1$$

$$Y_{\text{pred}} = Y_{\text{exp}} + \Delta Y_1 \tag{6.4}$$

where ΔX_1 and ΔY_1 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

$$X_{\text{pred}} = X_{\text{exp}} + \Delta X_3$$

$$Y_{\text{pred}} = Y_{\text{exp}} + \Delta Y_3 \tag{6.5}$$

and so on.

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

$$X_{\text{pred}} = X_{\text{exp}} + \Delta X_2$$

where

$$X_2 = (\Delta X_1 + \Delta X_3 + \Delta X_6)/3$$

$$X_{\text{pred}} = Y_{\text{exp}} + \Delta Y_2$$

and

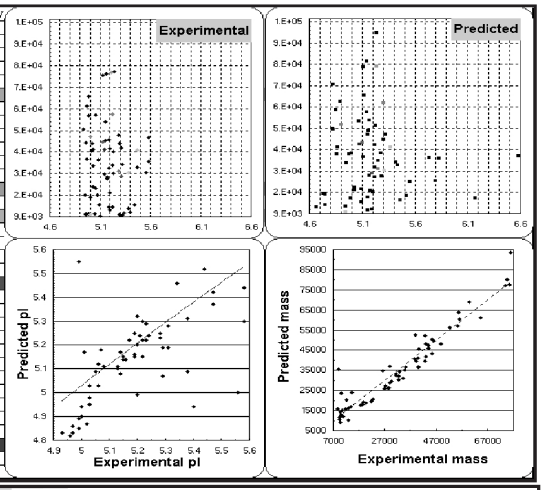
$$\Delta Y_2 = (\Delta Y_1 + \Delta Y_3 + \Delta Y_6)/3 \tag{6.6}$$

The outcome of this two-dimensional 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[16–18] 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.[19] Figure 6.8 depicts the overlap of observed and corresponding computed pI/Mw values for 40 proteins. A quantitative

(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	43089	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	P05125	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	35572	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

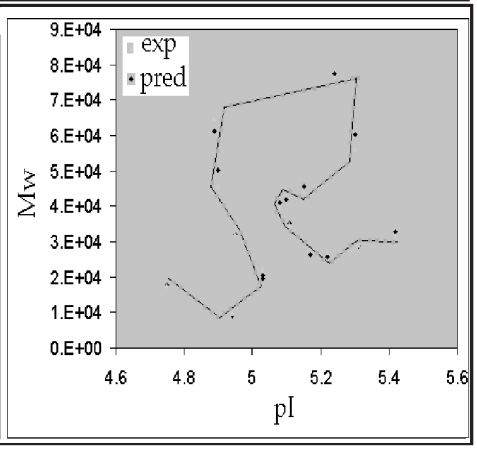


FIGURE 6.8 (a) Comparison of the values of isoelectric focusing points and molecular mass extracted from a high-resolution *E. coli* 2D PAGE map downloaded from Swiss-2D PAGE and those computed in this work. In the two upper charts, a small number of corresponding data points from each set have the same color for a quicker visual inspection. (b) For a small subset of proteins, computed pI/MW values are fairly close to the experimental counterparts, providing a "constellation" of reference points that can be used for warping.

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.

$$r.s = [(\Delta pI/pI_{exp})^2 + (\Delta Mw/Mw_{exp})^2]^{1/2}$$

where

$$\Delta pI = pI_{exp} - pI_{pred} \text{ and } \Delta Mw = Mw_{exp} - Mw_{pred} \quad (6.7)$$

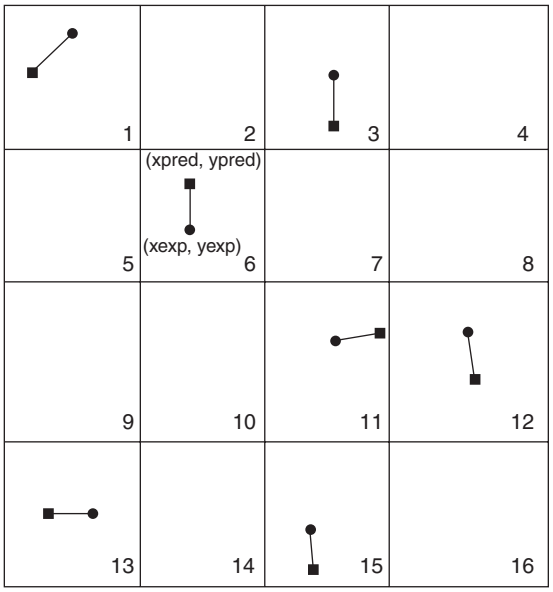


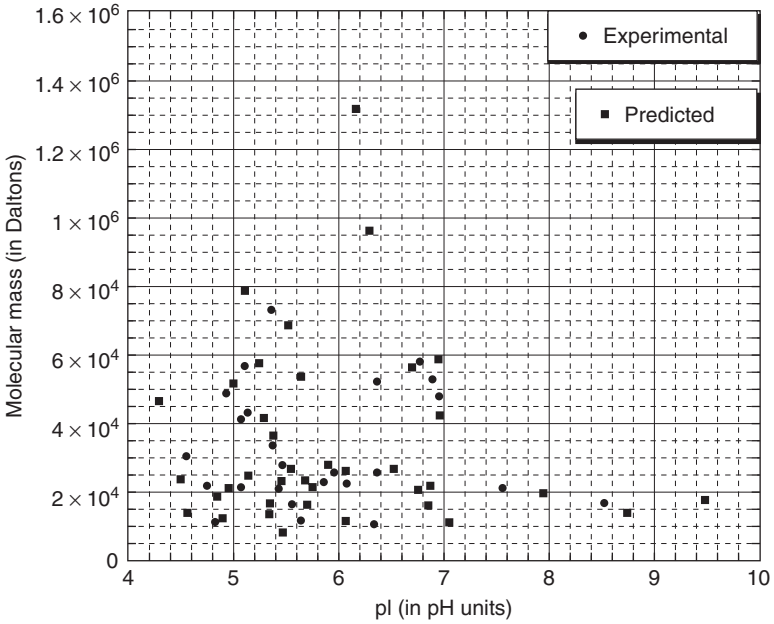
FIGURE 6.9 The warping of a 2D PAGE map on a computed pI/MW chart can be achieved by dividing it in areas surrounding each pair of experimental (●) and predicted (■) landmarks and applying to all the protein spots belonging in a particular neighborhood the necessary local translation to transform the coordinates (Xpred, Ypred) to (Xexp, Yexp).

Au: Symbols?

Despite the broad nominal intervals for pI (4-8 pH units) and MW (0-200kD), 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 types of modifications occurring co- and post-translationally that can severely alter the electrophoretic mobility of the proteins affected. As can be seen in Figure 6.9, 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 pI/Mw data points onto a set of experimentally determined protein spots without individually identifying and incorporating the aforementioned corrections in the computation of these attributes.

6.3 TMAP (TISSUE MOLECULAR ANATOMY PROJECT)

By mining publicly accessible databases, we have developed a collection of tissue-specific predictive protein expression maps (PEM) as a function of cancer histological state. Data analysis is applied to the differential expression of gene products in pooled libraries from the normal to the altered state(s). We wish to report the initial results of our survey across different tissues and explore the extent to which this comparative approach may help uncover panels of potential biomarkers of tumorigenesis, which would warrant further examination in the laboratory. For the third



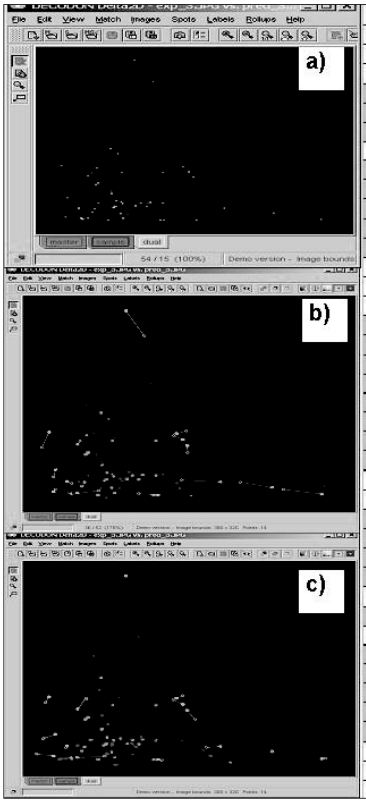
This is the same caption for Figure 8

FIGURE 6.10 (a) Comparison of the values of isoelectric focusing points and molecular mass extracted from a high-resolution *E. coli* 2D PAGE map downloaded from Swiss-2D 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 pI/MW values are fairly close to the experimentally counterparts, providing a "constellation" of reference points that can be used for warping.

dimension, we computed inferred gene-product translational expression levels from the transcriptional levels reported in the public databases. A number of studies[2,6], have explored the feasibility of molecular characterization of the histopathological state from the mRNA abundance reported in public databases. Many potential tissue-specific cancer biomarkers were tentatively identified as a result of mining expression databases. Thus arose the motivation to explore and catalogue correlations across different tissues as a first step toward comparative cancer proteomics of normal vs. diseased state. One potential clinical application is uncovering threads of biomarkers and therapeutic targets for multiple cancers.

6.3.1 Data Mining

For each tissue, the CGAP database can be queried by possible histological state, source, extraction, and cloning method. In the initial construction of queries, selecting the option "ANY" from within all of these fields provides an initial overview of the available libraries available. The more restrictive the search, the fewer libraries were selected. Within each library, transcripts are listed along with the number of times they were detected after a fixed number of PCR cycles. Since we were primarily



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	21019	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
5.06	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	11038	3.48	17819	0.997
5.31	11038	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	26629	0.039
5.56	16289	5.7	15805	0.039
5.59	52522	5.61	54265	0.033
5.6	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	0.039
5.95	25634	5.88	28355	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	6.95	59625	0.043
6.89	52776	6.71	56008	0.067
6.96	48310	6.97	42769	0.115
7.57	21182	6.86	22204	0.105
8.52	16844	7.93	19874	0.188
8.52	16844	6.76	20159	0.285
9.44	11331	8.73	15126	0.343

FIGURE 6.11 (a) Overlap of spots identified in 2D PAGE map of human colorectal epithelial cell line (in green) and theoretically computed (in 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 posttranslation modifications are simply too large to successfully align. This analysis was carried out using a demonstration version of the Delta-2D package.[18]

interested in computing protein maps, the gene symbols associated with those ESTs that were clustered to a gene of known function were extracted from UNIGENE. A Perl script performed the cross-reference checking between the two data sets and output a list of gene symbols and corresponding Swiss-Prot/trEMBL accession numbers (AC). The list of resulting AC was input to the pI/Mw tool server, which computed the necessary pI (isoelectric focusing point) and molecular mass (Mw) for the mature, unmodified proteins.[12] In the case of a single library, this information was married to the expression-detection counts in the following manner: The number of hits for each EST was first divided by the sum total of sequences within that library to provide a relative expression for each transcript. Finally, a renormalization was carried out by dividing relative expression levels by the maximum relative

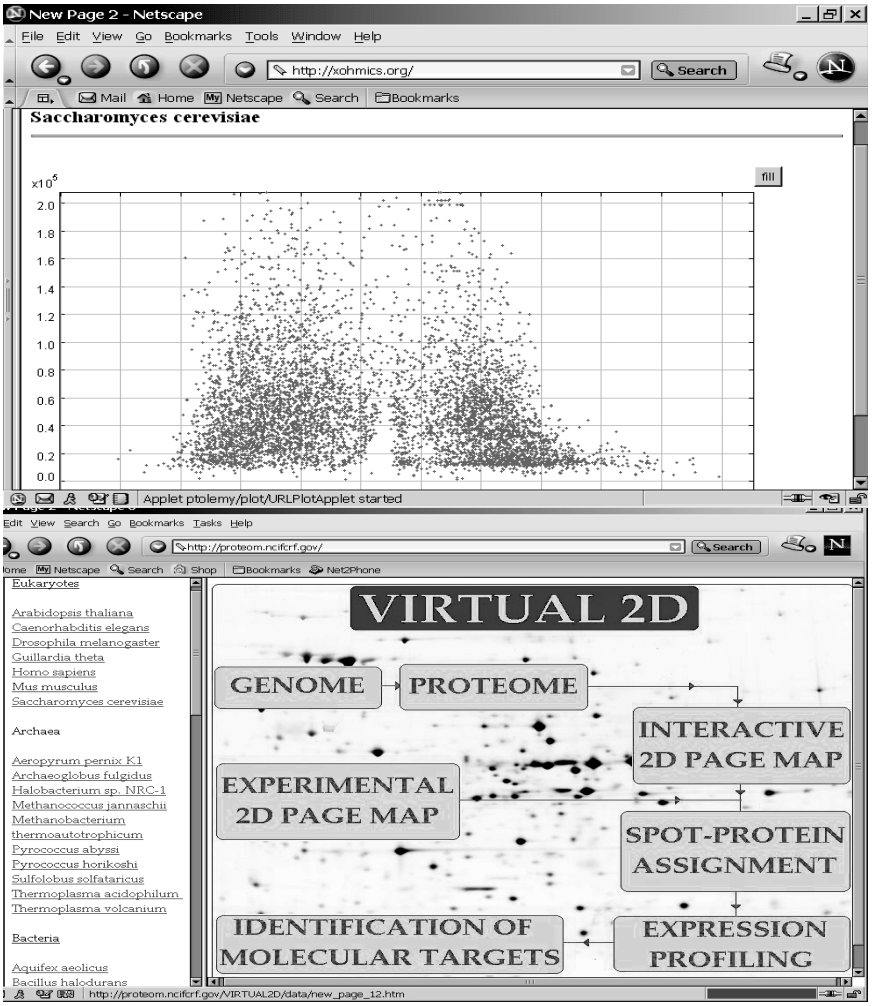


FIGURE 6.12 A snapshot of the screen display of VIRTUAL2D protein expression maps computed for ninety-two organisms/proteomes using data obtained from the European Bioinformatics Institute[24] can be displayed by clicking on any of the entries on the left.

expression level. In the event that a tissue search revealed several libraries fulfilling the requirements of the initial query, to improve the signal-to-noise ratio, the results are first pooled to generate a nonredundant list of entries and a more comprehensive expression map for that tissue and corresponding to that histological state. ~~The flow chart is depicted in Figure 6.13~~.

6.3.2 PROTPLOT

ProtPlot is a Java-based data-mining software tool for virtual 2D gels. It was derived from Opensource MAExplorer project (MAExplorer.sourceforge.net). It may be

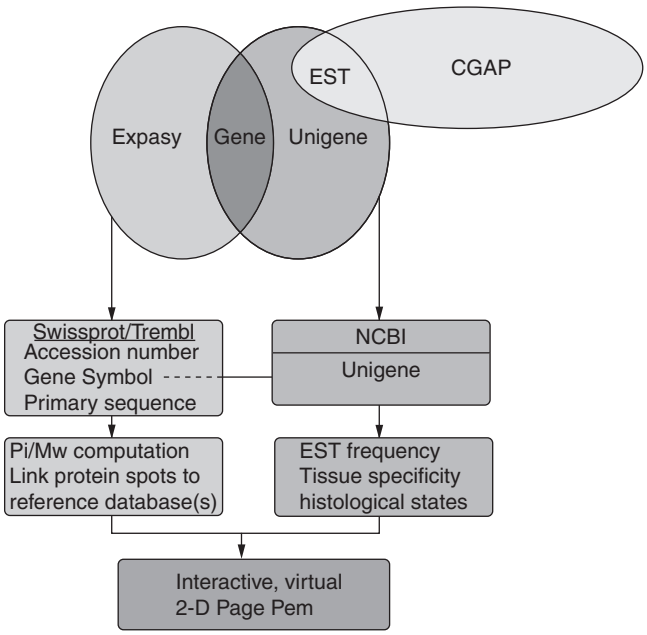


FIGURE 6.13 Overview of the public databases used and mining strategy.

downloaded and run as a stand-alone application. Its exploratory data analysis environment provides tools for the data mining of quantified virtual 2D gel (pIe, Mw, expression) data of estimated expression from the CGAP EST mRNA tissue expression database. This lets one look at the aggregated data in new ways; for example, which estimated "proteins" are in a specified range of (pI,Mw)? Or which sets of estimated "proteins" are up- or downregulated or missing between cancer samples and normal samples? Which sets of "proteins" cluster together across different types of cancers or normals? Here, one may aggregate several different normal and several different cancers as well as specify other filtering criteria.

As is well known, mRNA expression generally does not correlate well with protein expression as seen in 2D PAGE gels.[20] However, some new insights may occur by viewing the transcription data in the protein domain. If actual protein expression data is available for some of these tissues, it might be useful to compare mRNA estimated expression and actual protein expression. This tool may help find those proteins with similar expression and those that have quite different expression. This might be useful in thinking about new hypotheses for protein post-modifications or mRNA posttranscription processing.

ProtPlot generates an interactive virtual protein 2D gel Map scatterplot based on a database of derived maximum EST expression over a variety of tissue types from data obtained from the NCI-NCBI CGAP EST database of human cancer, precancer, and cancer mRNA expression (CGAP is the NCI's Cancer Genome Anatomy Project [http://cgap.nci.nih.gov/]. EST is the expressed sequence tag of

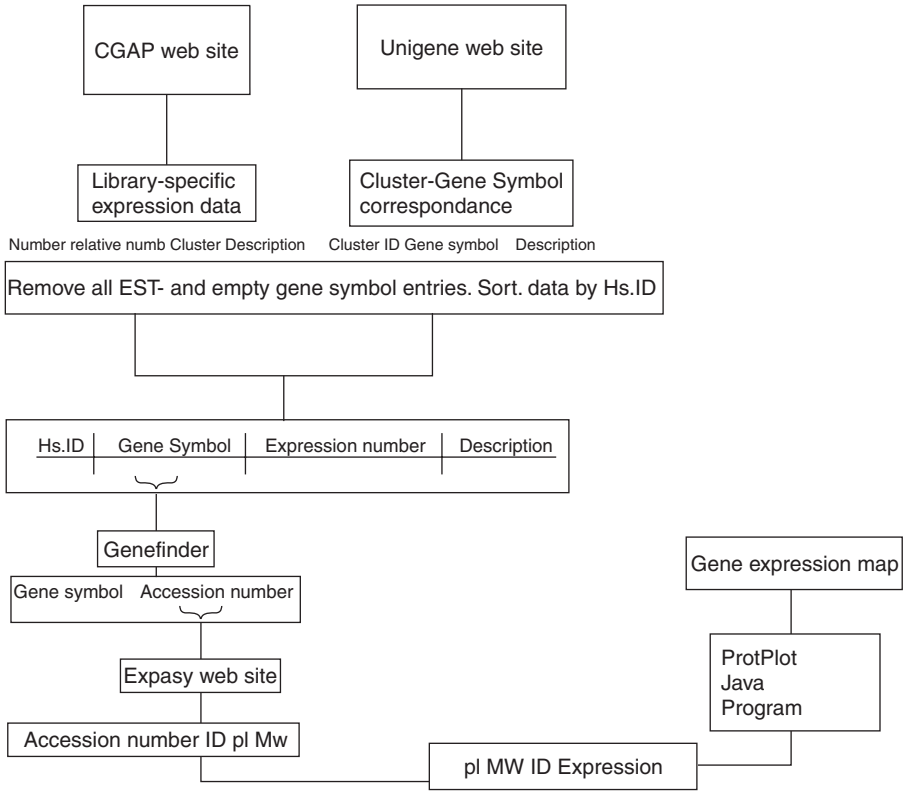


FIGURE 6.14 Flow chart describing in detail steps in the computation of expression maps.

mRNA found in particular tissues). The EST hit rate is a rough estimate of gene expression. These ESTs were mapped to Swiss-Prot (http://www.expasy.ch) accession numbers and Ids; the Mw and pI estimates were computed and used as estimates for corresponding proteins in a pseudo 2D gel.

ProtPlot data is contained in a set of tissue- and histology-specific .prp (i.e., ProtPlot) files described in the data format documentation. These are kept in the PRP directory that comes with ProtPlot when you install it. You will be able to update these .prp files from the ProtPlot Web server http://www.lecb.ncifcrf.gov/TMAP.

6.3.2.1 Using ProtPlot for Data Mining Virtual Protein Expression Patterns

First, one needs to download and install ProtPlot on a local computer. The detailed steps are shown in the following screen shots. This downloads the ProtPlot Java program and the CGAP-derived data set of pseudo 2D gels. If one downloads the

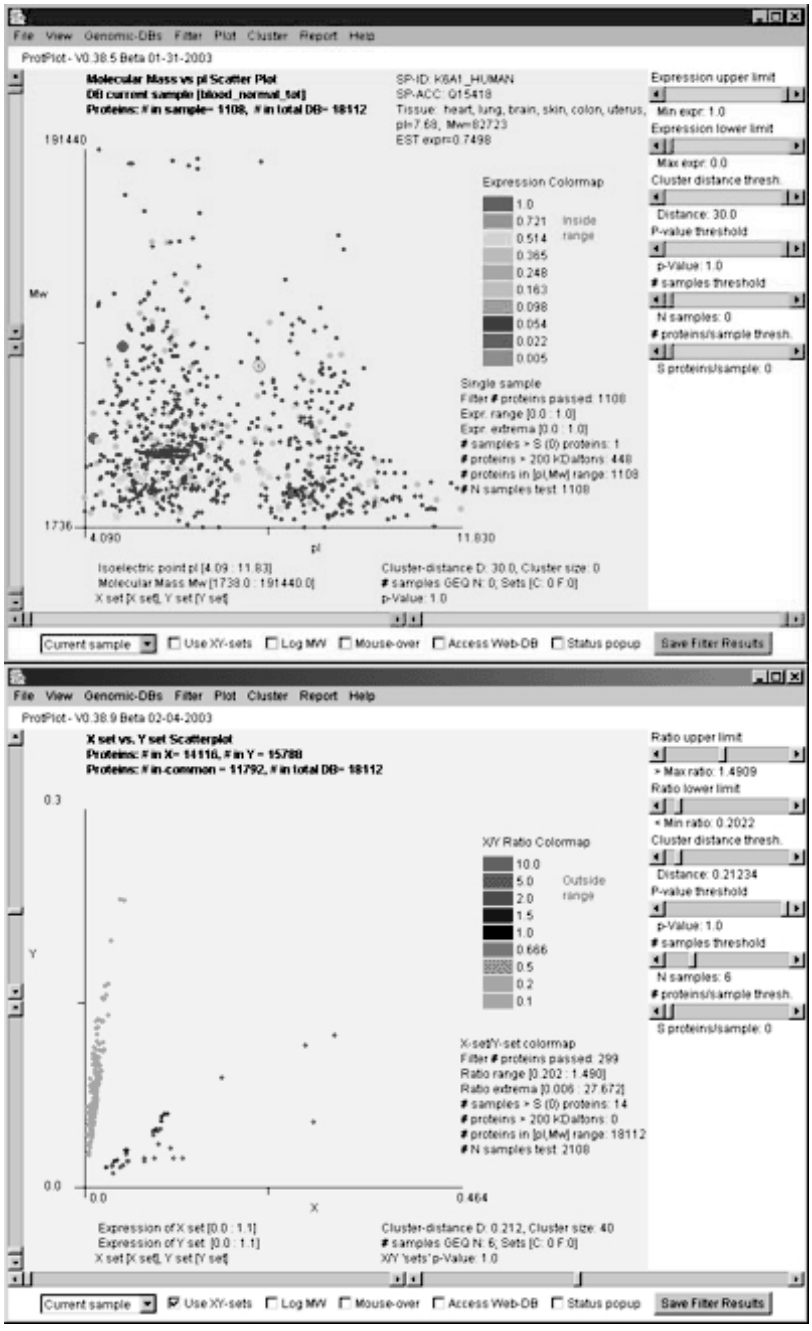
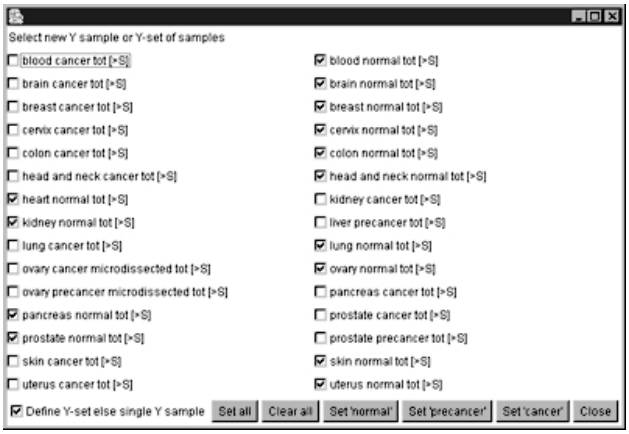


FIGURE 6.15 Snapshot of scatter plots from one sample in ProtPlot (Top). It is also possible to create (Bottom) an (X vs. Y) scatterplot or (mean X set vs. mean Y set) scatterplot when the corresponding ratio display mode is set. The following window shows the (mean X set vs. mean Y set) scatterplot.

(a)



(b)

FIGURE 6.16 (a) Tissue and histology selection panel. (b) This may be invoked either from the File menu or the pull-down sample selector at the lower left corner of the main window.

version that includes the Java Virtual Machine (JVM), it will not interact with any other JVM installed.

ProtPlot is started by clicking on the ProtPlot startup icon (Windows, MacOS-X, etc.) or by typing ProtPlot on the command line (Unix, Linux, and other systems).

Once the ProtPlot program is started, it loads the set of PRP files that were downloaded with the ProtPlot program. The virtual protein data for each tissue is used to construct a master protein index where proteins will be present for some tissues and not for others. The data is presented in a pseudo 2D gel image with the estimated isoelectric point (pI) on the horizontal axis and the molecular mass (Mw)

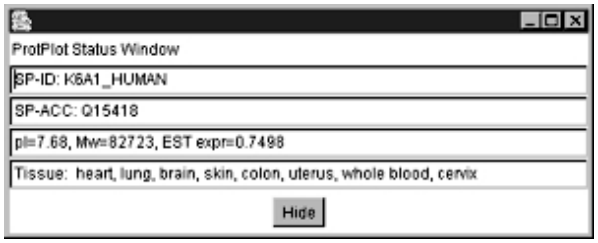


FIGURE 6.17 Snapshot of popup status window.

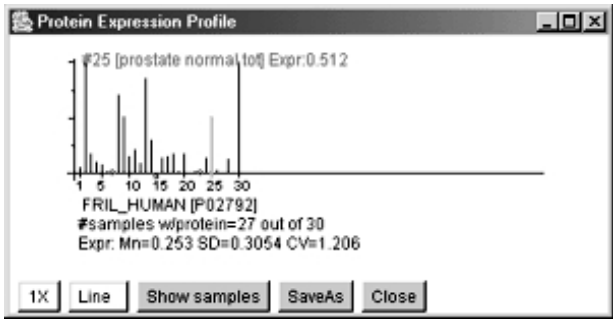


FIGURE 6.18 One can at glance obtain the expression profile of proteins or groups of proteins across tissues of choice.

on the vertical axis. Sliders on each of the axes allow one control the minimum and maximum values of pI and Mw displayed and thus the Mw vs. pI scatterplot zoom region one wants to select. By clicking on a spot in the scatterplot, the information on that protein will be displayed. One can also define that protein as the current protein. The current protein is used in some of the clustering methods, protein-specific reports (expression profile report), and the expression profile plot. If one has enabled the popup Genomic-ID web browser and is connected to the Internet, a web page from the selected Genomic database for that protein will pop up. One then selects various options from the pull-down menus. Some of the more commonly used options are replicated as check boxes at the bottom of the window.

6.3.2.2 The Scatterplot Display Mode

There are two primary types of pseudo 2D gel (Mw vs. pI) scatterplot display modes (summarized in Table 6.2) of this derived protein expression data: expression mode or ratio mode. The expression data may be for a single sample (the current sample) or the mean expression of a list of samples (called the expression profile, or EP). The ratio data is computed as the ratio of two individual samples called *X* and *Y*.

TABLE 6.2

Display Mode	Current Sample	Single X/Y	X Set/Y Set	EP Set
Expression	Yes	No	No	No
Single samples ratio	No	Yes	No	No
X-set and Y-set samples ratio	No	No	Yes	No
Mean Expression	No	No	No	Yes

TABLE 6.3

Filter Name	Current Sample	Single *X*/*Y*	*X* Set/*Y* Set	EP Set
> 200K Daltons	Yes	Yes	Yes	Yes
Tissue type	Yes	Yes	Yes	Yes
Expression (Ratio) range	expression	Ratio	Ration	expression
X/*Y* (inside/outside) range	No	Yes	Yes	No
(*X* set, *Y* set) *t*-Test	No	Yes	Yes	No
(*X* set, *Y* set) KS—Test	No	Yes	Yes	No
(*X* set, *Y* set) Missing data	No	Yes	Yes	No
At Most (Least) N samples	No	No	Yes	Yes
AND of saved cluster set	Yes	Yes	Yes	Yes
AND of saved filter set	Yes	Yes	Yes	Yes

Ratio data may alternatively be computed from sets of *X* samples and sets of *Y* samples. Generally, one would group a set of samples with similar characteristics together having the same condition (e.g., cancer, normal, etc.). The ratio of *X* and *Y* may be single samples, in which case the ratio is computed as:

$$\text{Ratio} = (\text{expression } X/\text{expression } Y) \tag{6.9}$$

where expression *X*/(expression *Y*) is the expression of corresponding proteins. Alternatively, one may compute the ratio of the mean expression of two different sets of samples (the *X* set and the *Y* set). The *X* and *Y* sets may be thought of as experimental conditions and the members of the sets being "replicates" in some sense; e.g., the *X* set could be cancer samples and the *Y* set could be normal samples. The ratio of the *X*/*Y* sets for each corresponding protein is computed as:

$$\text{Ratio} = (\text{mean } X - \text{set expression}/\text{mean } Y - \text{set expression}) \tag{6.10}$$

~~The following~~ shows a screen shot of one of the (Mw vs. pI) scatterplots when the display mode was set to (*X* set/*Y* set) ratio mode.

6.3.2.3 Effect of Display Mode on Filtering, Clustering, and Reporting

A particular display mode is selected using the Plot menu commands. When one selects a particular display mode, it will enable and disable Filter, View, Cluster, and Report options depending on the mode. For example, one may only use the *t*-test or missing *XY* set test if one is in *XY* sets ratio mode. Clustering can only be performed in EP set mode. One may change the display mode using the Plot menu | Show Display mode commands. Alternatively, since it is used so often, there is a

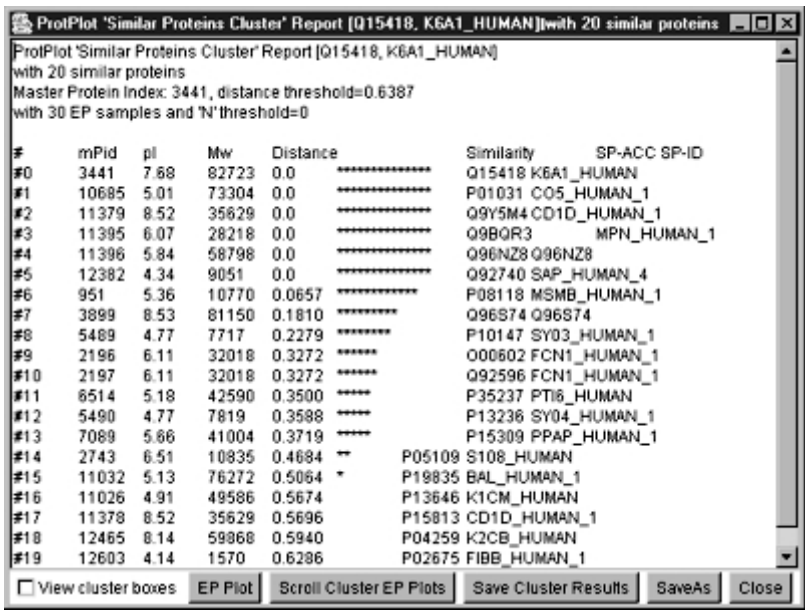


FIGURE 6.19 This window illustrates the scrollable list of EP plots sorted by the current cluster report similarity.

check box at the bottom of the main window "Use *XY* sets" that will toggle between the *XY* sets ratio mode and the previously set mode.

6.3.2.4 Selecting Samples

Samples for the current sample, *X* sample, *Y* sample, *X* set samples, *Y* set samples, and EP-set samples are selected using a popup check box list chooser of all samples. For example, one may invoke this chooser for the specific tissue sample one wants to view by using the File menu | Select samples | Select Current PRP sample. For *X*(*Y*) data, one invoke the choosers using File menu | Select samples | Select *X*(*Y*) PRP sample(s). One may switch between single (*X*/*Y*) and (*X* set/*Y* set) mode using the File menu | Select samples | Use Sample *X* and *Y* sets else single *X* and *Y* samples (CB) command.

There is an alternative display called the Expression Profile (EP) plot, which display a list of a subset of PRP samples for the currently selected protein. One may also display the scatterplot on the mean EP data for all proteins. The EP samples are specified using the File menu | Select samples | Select Expression List of samples command.

6.3.2.5 Listing a Report on Sample Assignments

A report of the current sample assignments for the current sample single *X* sample, single *Y* sample, *X* sample set, *Y* sample set, and EP sample set may be obtained using the File menu | Select samples | List sample assignments command.

6.3.2.6 Assigning the *X* Set and *Y* Set Condition Names

The default experimental condition names for the *X* and *Y* sample sets are "*X* set" and "*Y* set." One may change these by the File menu | Select samples | Assign *X* (*Y*) set name commands.

6.3.2.7 Status Reporting Window

There is a status popup window that first appears when the program is started and reports the progress while the data is loading. After the data is loaded, it will disappear. Toggling the "Status popup" checkbox at the bottom of the window will make it reappear. One may also press the "Hide" button on the status popup window to make it disappear.

6.3.2.8 Data Filtering

The pseudo-protein data is passed through a data filter consisting of the intersection of several tests including: pI range, MW range, sample expression range, expression ratio (*X*/*Y*) range (either inside or outside the range), *t*-test comparing the *X* and *Y* sample sets, Kolmogorov-Smirnov test comparing the *X* and *Y* sample sets, missing proteins test for *X* and *Y* sample sets, tissue type filter, protein family filter (to be implemented), and clustering. The filtering options are selected in the Filter menu. Looking at the scatterplot in ratio mode, one may filter by ratio of *X*/*Y* either inside or outside of the ratio range. The missing protein test defines "missing" as totally missing and "present" as having at least "N" samples present. Note that the *t*-test and the missing protein test are mutually exclusive in what they are looking for, so using both results in no proteins found.

6.3.2.9 Saving Filtered Proteins in Sets for Use in Subsequent Data Filtering

One may save the set of proteins created by the current data filter settings by pressing the "Save Filter Results" button in the lower right of the main window. This set of proteins is available for use in future data filtering using the Filter menu | Filter by AND of Saved Filter proteins (CB). Saving the state of the ProtPlot database (Filter menu | State | Save State) will also write out the save protein sets (saved filtered proteins and saved clustered proteins) in the database "Set" folder with ".set" file name extensions. In the Filter menu | State | Protein Sets submenu there are a number of commands to manipulate protein set files. One may individually save (or restore) any particular saved filtered set to (or from) a set file in the Set folder. There are also commands to compute the set intersection, union, or difference between two protein set files and leave the resulting protein set in the saved Filter set.

6.3.2.10 Filter Dependence on the Display Mode

Au: Which table?

Note that the particular filter options available at any time depend on what the current display mode is. Table shows which options are available for which display modes.

TABLE 6.4

Filter Name	Current Sample	Single *X*/*Y*	*X* Set/*Y* Set	EP Set
Statistics or proteins passing filter	SP-ACC/ID, pI, Mw, expression	SP-ACC/ID, pI, Mw, *X*/*Y*, *X*, *Y* expr, tissues	SP-ACC/ID, pI, Mw, mn*X*/mn*Y*, (mn, sd, cv, n) expr. for *X* and *Y* sets, tissues. If using *t*-test then (dF, *t*-stat, *F*-stat). If using KS-test then (dF, D-stat)	SP-ACC/ID, pI, Mw, (mn, sd, cv, n) expr.c for EP set, tissues
Expression profiles of proteins passing filter	SP-ACC/ID, expr. data EP set	SP-ACC/ID, expr data EP set	SP-ACC/ID, expr. data EP set	SP-ACC/ID, expr. data EP set
X and *Y* sets of missing proteins pasing filter	No	No	SP-ACC/ID, (mn, sd, cv, n) for *X* and *Y* sets	No
EP set statistics of proteins passing filter	No	No	No	SP-ACC/ID, (mn, sd, cv, n) for EP set
List of samples in current EP profile	{Nbr., sample-name, expression)	{Nbr, sample-name, expression)	{Nbr, sample-name, expression)	{Nbr., sample-name, expression)
List of all sample assignments	Current, *X*, *Y*, *X* set, *Y* set, EP set	Current, *X*, *Y*, *X* set, *Y* set, EP set	Current, *X*, *Y*, *X* set, *Y* set, EP set	Current, *X*, *Y*, *X* set, *Y* set, EP set
List of # proteins/sample	{Sample-name, # proteins in sample}	{Sample-name, # proteins in sample}	{Sample-name, # proteins in sample}	{Sample-name, # proteins in sample}
ProtPlot state	State	State	State	State

6.3.2.11 The Data Mining State

The current data mining settings of ProtPlot are called the "state." They may be saved in a named startup file called the "startup state file" in the State folder. The State folder and other folders used by ProtPlot are found in the directory where ProtPlot is installed. Initially there is no startup state file. If one saves the state, then this file is created. As many of these saved state files can be created as desired. One may change the file and thus save various combinations of settings of samples for the current, *X*, *Y*, and expression list of samples. The state also includes the various filter, view, and plot options as well as the pI, Mw, expression, ratio, cluster distance threshold, number samples threshold, *p* value threshold sliders, and other settings. The saved Filter and Cluster sets of proteins are also written out as .set files in the Set folder when the state was saved.

Starting ProtPlot by clicking on the ProtPlot startup icon will not read the state file when it starts up. However, if a state is saved, clicking on the state file or a shortcut to the state file will cause it to be read when ProtPlot starts up.

The current state can be saved using either the File | State | Save State command to save it under the current name or the File | State | Save As State command to save it under a new name. The current state may be changed using File | State | Open State file command.

6.3.2.12 The Molecular Mass vs. pI Scatterplot: Expression or Ratio

There are two types of scatterplots: expression for a single sample or the ratio of two samples *X* and *Y*. The Plot menu lets one switch the display mode. Ratio mode itself has two types of displays: red (*X*) + green (*Y*), or a ratio scale ranging between <1/10 (green) and >10 (red). One may view a popup report of the expression or ratio values for the current protein. If "mouse-over" is enabled, then moving the mouse over a spot will show the name of the protein and its associated data. If mouse-over is not enabled, then clicking on the spot will show its associated data. One may scroll the scatterplot in both the pI and Mw axes by adjusting the endpoint scrollbars on the corresponding axes. In addition, one may display the scatterplot with a log transform of MW by toggling the log MW switch.

The popup plots and scatterplot may be saved as .gif image files, which are put into the project's Report folder. Similarly, reports are saved as tab-delimited .txt text files in the Report folder. Because a file name is prompted for, one may browse one's file system and save the file in another disk location.

6.3.2.13 *X* Sample(s) vs. *Y* Samples Scatterplot

In *X*/*Y* ratio mode (single *X*/*Y* samples or *X*-set/*Y*-set samples), a scatterplot of the *X* vs *Y* expression data can be viewed. Enable the *XY* scatterplot using the Plot menu | Display (*X* vs. *Y*) else (Mw vs pI) scatterplot if ratio mode (CB). The scatterplot can be zoomed similar to the Mw vs. pI scatterplot. The proteins displayed are those passing the data filter that have both *X* and *Y* data (i.e., expression is > 0.0).

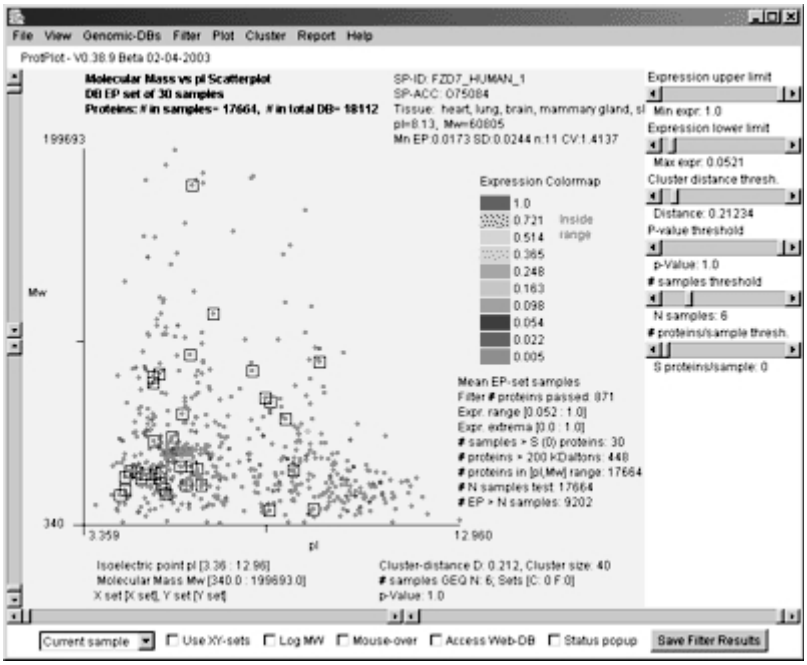


FIGURE 6.20 The spots marked by boxes belong to the same cluster.

6.3.2.14 Expression Profile Plot of a Specific Protein

An expression profile (EP) shows the expression for a particular protein for all samples that have that protein. The Plot menu | Enable expression profile plot pops up an EP plot window and displays the EP plot for any protein selected. The relative expression is on the vertical axis and the sample number on the horizontal axis. Pressing on the "Show samples" button pops up a list showing the samples and their order in the plot. Pressing on the "n×" button will toggle through a range of magnifications from 1× through 50× that may be useful in visualizing low values of expression. Clicking on a new spot in the Mw vs. pI scatterplot will change the protein being displayed in the EP plot. Within the EP plot display, one may display the sample and expression value for a plotted bar by clicking on the bar (which changes to green with the value in red at the top). The EP plot can be saved as a .gif file. One may also click on the display to find out the value and sample. Note: since clustering uses the expression profile, one must be in "mean EP-set display" mode.

6.3.2.15 Clustering of Expression Profiles

One may cluster proteins by the similarity of their expression profiles. First set the plot display mode to "Show mean EP-set samples expression data." The clustering method is selected from the Cluster menu. Currently there is one cluster method; others are planned. The cluster distance metric is the distance between two proteins

based on their expression profile. The metric may be selected in the Cluster menu. Currently, there is one clustering method: cluster proteins most similar to the current protein (specified by clicking on a spot in the scatterplot or using the Find Protein by name in the Files menu). It requires one to specify a) the current protein, and b) the threshold distance cutoff. The threshold distance is specified interactively by the "Distance Threshold T" slider. The Similar Proteins Cluster Report will be updated if either the current protein or the cluster distance is changed.

The cluster distance metric must be computed in a way to take missing data into account since a simple Eucledian distance cannot be used with the type of sparse data present in the ProtPlot database. ProtPlot has several ways to compute the distance metric using various models for handling missing data.

One may save the set of proteins created by the current clustering settings by pressing the Save Cluster Results button in the lower right of the cluster report window. This set of proteins is available for use in future data filtering using the Filter menu | Filter by AND of Saved Clustered proteins (CB). When the state of the ProtPlot database is saved (Filter menu | State | Save State), the set of saved clustered proteins will be saved in the database Set folder. One may restore any particular saved clustered set file.

The EP plot window can be brought up by clicking on the EP Plot button and then click on any spot in the scatterplot to see its expression profile. Clicking on the Scroll Cluster EP Plots button brings up a scrollable list of expression profiles for just the clustered proteins sorted by similarity.

One may mark the proteins belonging to the cluster in the scatterplot with black boxes by selecting the View Cluster Boxes check box at the lower left of the cluster report window. This is illustrated in the ~~following window~~:

Au: is this a figure?

6.3.2.16 Reports

Various popup report summaries are available depending on the display mode. All reports are tab-delimited and so may be cut and pasted into MS Excel or other analysis software. Reports also have a "Save As" button so data can be saved into a tab-delimited file. The default/Report directory is in the directory where ProtPlot is installed. However, it can be saved anywhere on one's file system. The content of some reports depends on the particular display mode. This is summarized in the table below.

Au: Which table?

6.3.2.17 Genomic Databases

If one is connected to the Internet and have enabled ProtPlot to "Access Web-DB," then clicking on a protein will popup a genomic database entry for that protein. The particular genomic database to use is selected in the Genomic-DB menu.

6.4 RESULTS AND DATA ANALYSIS

Figure 6.21 depicts the pI/Mw maps computed by our approach for a number of these tissues. They all display the characteristic bimodal distribution that was explained previously as the statistical outcome of a limited, pK-segregated proteomic

TABLE 6.5

Blood		Brain		Breast	
Upregulated	**Downregulated**	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	O00215	P04075	O00184	P02571	O43443
	P01907	P12277	O14498	P05388	O43444
	P01909	P41134	O15090	P12751	O60930
	P05120	P15880	O95360	P18084	O75574
	P35221	P12751	P01116	P49447	P15880
	P42704	P02570	P01118	Q05472	P17535
	P55884	P70514	P02096	Q15445	P19367
	Q29882	P99021	P20810	Q9BTP3	Q96HC8
	Q29890	Q11211	P50876	Q9HBV7	Q96PJ2
	Q99613	P46783	Q9BZZ7	Q9NZH7	Q96PJ6
	Q99848	P26373	Q9UM54	Q9UBQ5	Q9NNZ4
	Q9BD37	P26641	Q9Y6Z7	Q9UJT3	Q9NNZ5

Cervix		Colon		Head and Neck	
Upregulated	**Downregulated**	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	O75331	P00354	O14732	O75770	O60573
	O75352	P02571	P00746	P00354	O60629
	P09234	P04406	P09497	P04406	O75349
	P11216	P04687	P17066	P06702	P30499
	P13646	P04720	P18065	P09211	P35237
	P28072	P04765	P38663	P10321	P49207
	P47914	P09651	P41240	P21741	P82909
	Q02543	P11940	P53365	P30509	Q9BUZ2
	Q9NPX8	P17861	P54259	Q01469	Q9H2H4
	Q9UBR2	P26641	Q12968	Q92597	Q9H5U0
	Q9UQV5	P39019	Q9P1X1	Q9NQ38	Q9UHZ1
	Q9UQV6	P39023	Q9P2R8	Q9UBC9	Q9Y3U8

Kidney		Liver		Lung	
Upregulated	**Downregulated**	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
O43257	O60622	P11021	P02792	O95415	O60441
O43458	Q14442	P11518		P01860	O75918
O75243	Q8WX76	P19883		P50553	O75947
O75892	Q8WXP8	P21453		P98176	O95833
O76045	Q96T39	P35914		Q13045	P01160
Q15372	Q9H0T6	P36578		Q15764	P04270
Q969R3	Q9HBB5	P47914		Q92522	P05092
Q9BQZ7	Q9HBB6	Q05472		Q9BZL6	P05413
Q9BSN7	Q9HBB7	Q13609		Q9HBV7	P11016
Q9UIC2	Q9HBB8	Q969Z9		Q9NZH7	Q13563
Q9UPK7	Q9UK76	Q9BYY4		Q9UJT3	Q15816
Q9Y294	Q9UKI8	Q9NZM3		Q9UL69	Q16740

TABLE 6.5
Continued

Ovarian		Pancreas		Prostate	
Upregulated	**Downregulated**	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	P02461	P00338	P05451	O00141	O15228
	P02570	P02794	P15085	P08708	O43678
	P04792	P04720	P16233	P19013	P10909
	P07900	P05388	P17538	P48060	P11380
	P08865	P07339	P18621	Q01469	P11381
	P11142	P08865	P19835	Q01628	P98176
	P14678	P20908	P54317	Q01858	Q92522
	P16475	P26641	P55259	Q02295	Q92826
	P24572	P36578	Q92985	Q13740	Q99810
	Q15182	P39060	Q9NPH2	Q96HK8	Q9H1D6
	Q9UIS4	Q01130	Q9UIF1	Q96J15	Q9H1E3
	Q9UIS5	Q15094	Q9UL69	Q9C004	Q9H723

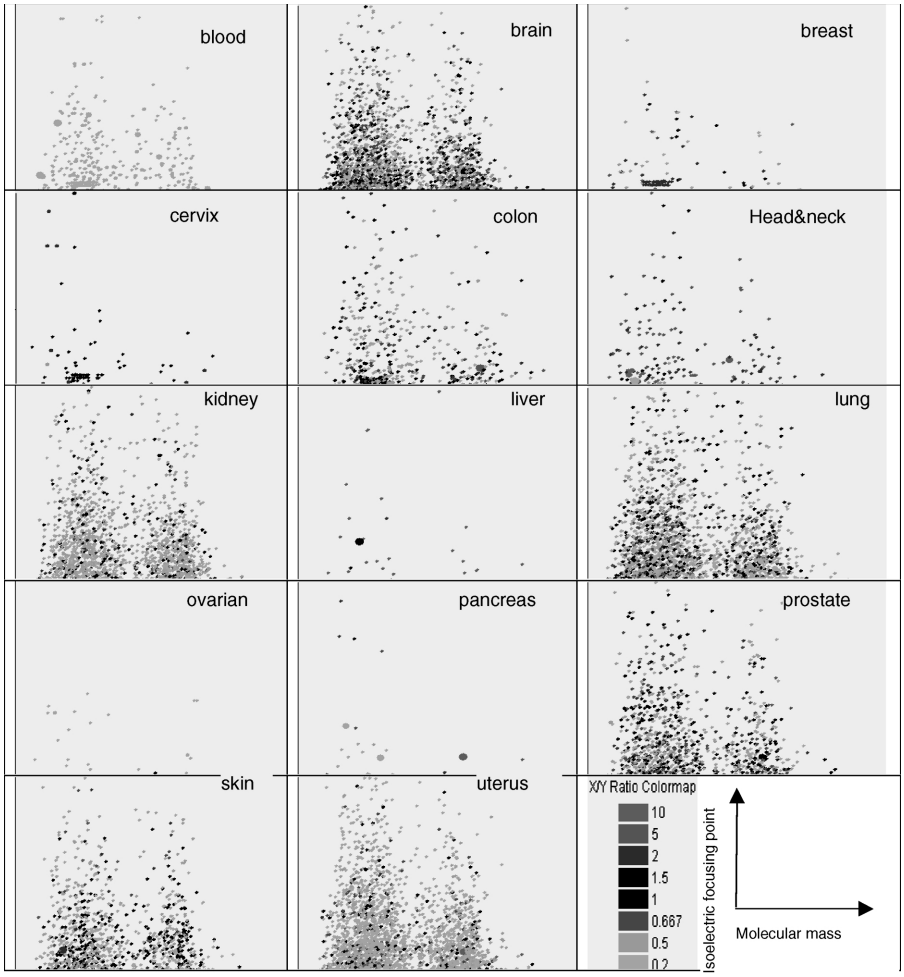
Skin		Uterus	
Upregulated	**Downregulated**	**Upregulated**	**Downregulated**
O14947	O00622		O95432
P01023	P12236		O95434
P02538	P12814		O95848
P06733	P19012		Q08371
Q02536	P28066		Q13219
Q02537	P30037		Q13642
Q13677	P30923		Q9UKZ8
Q13751	P33121		Q9UNK7
Q13752	P36222		Q9UQK1
Q13753	P43155		Q9Y627
Q14733	Q01581		Q9Y628
Q14941	Q9UID7		Q9Y630

alphabet.[12] In addition, one can quickly obtain the most significantly differentially expressed gene proteins by computing the tissue-specific charts of the ratios between normal and cancer states.

A number of proteins detected by the survey described are ribosomal or ribosomal-associated proteins (such as elongation factors P04720, P26641 in colon and pancreas). Their upregulation is consistent with an accelerated cancerous cell cycle. Others may turn out to be effective tissue-specific biomarkers such as phosphopyruvate hydratase (P06733 in skin). A third category will turn out to be druggable targets—molecular "switches" that can be the focus of drug design for therapeutic intervention to reverse or stop the disease.

Au: As meant?

However, identification of useful potential targets requires additional knowledge of their function and cellular location. Accessibility is an obvious advantage. Such is the case of laminin gamma-2 (Q13753), the second highest differentially expressed



Au: please provide caption.

FIGURE 6.21

protein in skin. It is thought to bind to cells via a high-affinity receptor and to mediate the attachment, migration, and organization of cells into tissues during embryonic development by interacting with other extracellular matrix components.

6.5 CONCLUSION

To date, the charts for 92 organisms have been assembled and are represented within VIRTUAL2D. TMAP results from the survey of 144 libraries from the CGAP public resource to produce more than 18,000 putative gene products encompassing normal, cancerous, and, when available, precancerous states for 14 tissues.

Au: "web-able" ok?

These interactive, web-~~able~~ knowledge based proteomics resources are available to the research community to generate and explore in the laboratory hypothesis-driven cancer biomarkers.

Au: please provide article titles for all journal references

REFERENCES

1. Fearon, E.R., Hamilton, S.R., and Volgeinstein, B. *Science,* 238, 193–197, 1987.
2. Page, M.J, Amess, B., Townsend, R.R., Parekh, R., Herath, A., Brusten, L., Zvelebil, M.J., Stein, R.C., Waterfield, M.D., Davies, S.C., and O'Hare, M.J. *Cell Biology,* 96, 22, 12589–12594, 1999.
3. Aebersold, R., Rist, B., and Gygi, S.P. *Ann NY Acad. Sci.,* 919, 33–47, 2000.
4. Bussow, K. *Trends Biotechnol.,* 19, 328–329, 2001.
5. Fivaz, M., Vilbois F., Pasquali, C., and van der Goot, F.G. *Electrophoresis,* 21, 3351–3356, 2000.
6. Kriegel, K., Seefeldt, I., Hoffmann, F., Schultz, C., Wenk, C., Regitz-Zagrosek, V., Oswald, H., and Fleck, E. *Electrophoresis,* 13, 2637–2640, 2000.
7. O'Farrell, P.H. *J. Biol. Chem.,* 250, 4007–4021, 1975.
8. O'Farrell, P.Z., Goodman, H.M., and O'Farrell, P.H. *Cell,* 12, 1133–1141, 1977.
9. Dihazi, H., Kessler, R., and Eschrich, K. *Anal. Biochem.,* 299, 260–263, 2001.
10. Angelis, F.D., Tullio, A.D., Spano, L., and Tucci, A.J. *Mass Spectrom.,* 36, 1241–1248, 2001.
11. Weiller, G.F., Djordjevic, M.J., Caraux, G., Chen, H., and Weinman, J.J. *Proteomics,* 12, 1489–1494, 2001.
12. Wulfkuhle, J.D., McLean, K.C., Paweletz, C.P., Sgroi, D.C., Trock, B.J., Steeg, P.S., and Petricoin, E.F., III. *Proteomics,* 10, 1205–1215, 2000.
13. Gorg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildruber, R., and Weiss, W. *Electrophoresis,* 6, 1037–1053, 2000.
14. Bjellqvist, B., Sanchez, J.C., Pasquali, C., Ravier, F., Paquet, N., Frutiger, S., Hughes, G.J., Hoschstrasser, and D.F. *Electrophoresis,* 14, 1375–1378, 1993.
15. VanBogelen, A.R., Abshire, Z.A., Moldover, B., Olson, R.E., and Neidhardt, C.F. *Electrophoresis,* 18, 1243–1251, 1197.
16. Lemkin, P.F., Thornwall, G., Walton, K., and Hennighausen, L. *Nucleic Acids Res.*, 22, 4452–4459, 2000.
17. Ptplot is a 2D data plotter and histogram tool implemented in Java that can be accessed at http://ptolemy.eecs.berkeley.edu/java/ptplot/
18. Information on Melanie (Geneva Bioinformatics) can be found at http://www.www.expasy.ch/melanie/.
19. Information about Z3 is available at http://www.2dgels.com/
20. Ideker et al., *Science,* 292, 929–934, 2001.
21. Bairoch, A. and Apweiler, R. *Nucleic Acids Res.,* 28, 45–48, 2000.
22. PI/MW is part of EXPASY's proteomics tools and can be accessed at http://www.expasy.ch/tools/pi_tool.html.
23. NCBI's Unigene database can be accessed at http://www.ncbi.nlm.nih.gov/UniGene.
24. The European Bioinformatics Institute web site can be found at http://www.ebi.ac.uk/

~~WORLD-2DPAGE. Index to 2-D PAGE databases and services http://www.expasy.ch/ch2d/2d-index.htm~~.

~~A typical case can be found at http://www.expasy.ch/cgibin/map2/def?HEPG2_HUMAN~~.

~~Information about SRS can be found at http://www.lionbioscience.com/solutions/srs~~.

Au: please cite all the following refs in text or delete. Will ~~need to~~ renumber ~~where inserted.~~

http://www-lecb.ncifcrf.gov/MAExplorer/

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, pp. 93–48.

Information on SPSS (Statistical Package for the Social Sciences) can be accessed at http://www.spss.com/.

Information on Delta2D from Decodon, Germany can be obtained by accessing their web site located at http://www.decodon.com/

Reymond, M.A., Sanchez, J. C., Hughes, G.J., Riese, J., Tortola, S., Peinado, M.A., Kirchner, T., Hohenberger, W., Hochstrasser, D.F., Kockerling, F., Electrophoresis, 1997; 18: 2842–2848.

Anderson, L., Seilhamer, J. Electrophoresis 1997; 18:11853–11861

Schuler, G.D. *J. Mol. Med.*, 1997; 75(10):694–698.

Boguski, M.S., Schuler, G.D., *Nat. Genetics,* 1995; 10:369–371.

An extensive body of information about the origin, statistics and current research methodologies in most commonly diagnosed types of cancers can be found at http://nci.nih.gov

Krizman, D.B., Chuaqui, R.F., Meltzer, P.S., Trent, J.M., Duray, P.H., Linehan, W.M., Liotta, L.A., and Emmert-Buck, M.R. *Cancer Res.,* 56, 5380–5383, 1996.

Strausberg, R.L., Dahl, C.A., Klausner, R.D. 1997, *Nat. Genetics,* 15, Spec, 415–416.

Information concerning the CGAP project can be found at http://egap.nci.nih.gov/.

Emmert-Buck, M.R., Bonner, R.F., Smith, P.D., Chuaqui, R.F., Zhuang, Z, Goldstein, S.R., Weiss, R.A., Liotta, L.A. Science, 1996, 8, 274(5289), 998–1000.

CGAP (Cancer genome anatomy projet). Additional information is available at http://CGAP.nci.nih.gov/

Grouse, L.H., Munson, P.J., Nelson, P.S., Urology, 2001, 57 (4 Suppl 1), 154–9.

Emmert-Buck, M.R., Strausberg, R.L., Krizman, D.B., Fatima Bonaldo, M., Bonner, R.F., Bostwick, D.G., Brown, M.R., Buetow, K.H., Chuaqui, R.F., Cole, K.A., Duray, P.H., Englert, C.R., Gillespie, J.W., Greenhut, S., Grouse, L., Hillier, L.W., Katz, K.S., Klausner, R.D., Kuznetzov, V., Lash, A.E., Lennon, G., Linehan, W.M., Liotta, L.A., Marra, M.A., Munson, P.J., Ornstein, D.K., Prabhu, V.V., Prange, C., Schuler, G.D., Soares, M.B., Tolstoshev, C.M., Vocke, C.D., Waterston, R.H., *Journal of Molecular Diagnostics,* 2000, 2, 60–66.

Loging, W.T., Lal, A., Siu, I.M., Loney, T.L., Wikstrand, C.J., Marra, M.A., Prange, C., Bigner, D.D., Strausberg, R.L., Riggins, G.J., *Genome Research,* 2000, 10, 1393–1402.

Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J., Altschul, S.F., *Genome Research,* 2000, 10:1051–1060.

Emmert-Buck, M.R., Strausberg, R.L., Krizman, D.B., Bonaldo, M.F., Bonner, R.F., Bostwick, D.G., Brown, M.R., Buetow, K.H., Chuaqui, R.F., Cole, K.A., Duray, P.H., Englert, C.R., Gillespie, J.W., Greenhut, S., Grouse, L., Hillier, L.W., Katz, K.S., Klausner, R.D., Kuznetzov, V., Lash, A.E., Lennon, G., Linehan, W.M., Liotta, L.A., Marra, M.A., Munson, P.J., Ornstein, D.K., Prabhu, V.V., Prange, C., Schuler, G.D., Soares, M.B., Tolstoshev, C.M., Vocke, C.D., Waterston, R.H. Molecular profiling of clinical tissue specimens: feasibility and applications. *Am J Pathol,* 2000, 156: 1109–1115.

Strausberg, R.L., Buetow, K.H., Emmert-Buck, M., Klausner, R. The Cancer Genome Anatomy Project: building an annotated gene index. *Trends in Genetics,* 2000, 16, 103–106.

Ryu, B., Jones, J., Hollingsworth, M.A., Hruban, R.H., Kern, S.E., *Cancer Research,* 2001, 61, 1833–8.

UNIGENE is a Database hosted by NCBI and can be accessed at http://www.ncbi.nlm.nih.gov/

SwissProt, a protein knowledgebase can be accessed at: http://www.expasy.ch

European Bioinformatics Institute.

Medjahed, D., Smythers, G., Stephens, M., Powell, D., Lemkin, P., Munroe, J. D. *Proteomics,* 2003, 2.

John, H. Gillespie, "Population Genetics: A Concise Guide," The Johns Hopkins University Press, Baltimore (1998) pg. 8.

Lee, J.-H., Kim, J.-M., Kim, M.-S., Lee, Y.-T., Marshak, D.R., Bae, Y.-S., *Biochem. Biophys. Res. Commun.*, 238:462–467(1997).

Lemkin, P.F., Thornwall, G.C., Walton, K.D., Hennighausen, L., Nucleic Acid Research, 2000, 28: 4452–4459. MAExplorer is available at http://maexplorer.sourceforge.net/