Djamel Medjahed[1,4]
Brian T. Luke[2,5]
Tawady S. Tontesh[1]
Gary W. Smythers[2,5]
David J. Munroe[1,4]
Peter F. Lemkin[3]

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

Tissue Molecular Anatomy Project (TMAP): An expression database for comparative cancer proteomics

By mining publicly accessible databases, we have developed a collection of tissue-specific predictive protein expression maps 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.

Keywords: Expressed sequence tags / Protein expression map / Tissue Molecular Anatomy Project / Two-dimensional gel electrophoresis / VIRTUAL2D PRO 0488

1 Introduction

The advent of the human genome sequence and subsequently of exquisitely curated proteome datasets provided the initial motivation to build interactive, web-able reference p*I*/M_r charts to facilitate the putative assignment of proteins to spots in experimental 2-D PAGE maps. The fruit of this effort known as VIRTUAL2D [1] has largely mirrored that of the European Bioinformatics Institute and has grown to comprise interactive web-able maps for ninety-two organisms and associated java-based graphics software. Molecular mass and isoelectric focusing points are not the only varying attributes predictable from the primary sequence of the gene product (*e.g.*, hydrophobicity, well characterized putative post-translational modifications, *etc.*).

For the third dimension, we computed inferred geneproduct translational expression levels from the transcriptional levels reported in the public databases. A number of studies [2, 3] 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 catalog correlations across different tissues as a first step towards comparative cancer proteomics of normal *versus* diseased state. One potential clinical application is uncovering threads of biomarkers and therapeutic targets for multiple cancers. For each expression sequence tag (EST), the number of hits detected in each tissue and state-specific library provides a third dimension, expression to these maps. They in turn can be thought of as snapshots of the profile at different states (drug-induced, diseased, *etc.*).

Cancer is a result of multiple pathways involving a succession of molecular alterations at the cellular level. These transformations may take place over an extended period of time and only after the cell has accumulated a critical number of these changes does it become cancerous. The Cancer Genome Anatomy Project (CGAP) [4] is a program launched by the US National Cancer Institute (NCI) to keep track of the gradual molecular alterations that occur throughout this transformation process. Not all genes are switched on for any given tissue. The expression profile can therefore be used to uniquely characterize each tissue, much like a fingerprint. Furthermore when a normal tissue is transformed into the cancerous state, the expression profile changes. These alterations in gene expression drastically affect the nature and amounts of proteins produced, disrupting their interaction networks. Many different scenarios of gene changes and protein interactions are observed in cancerous tissue.

2 Materials and methods

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

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

Abbreviation: CGAP, cancer genome anatomy project

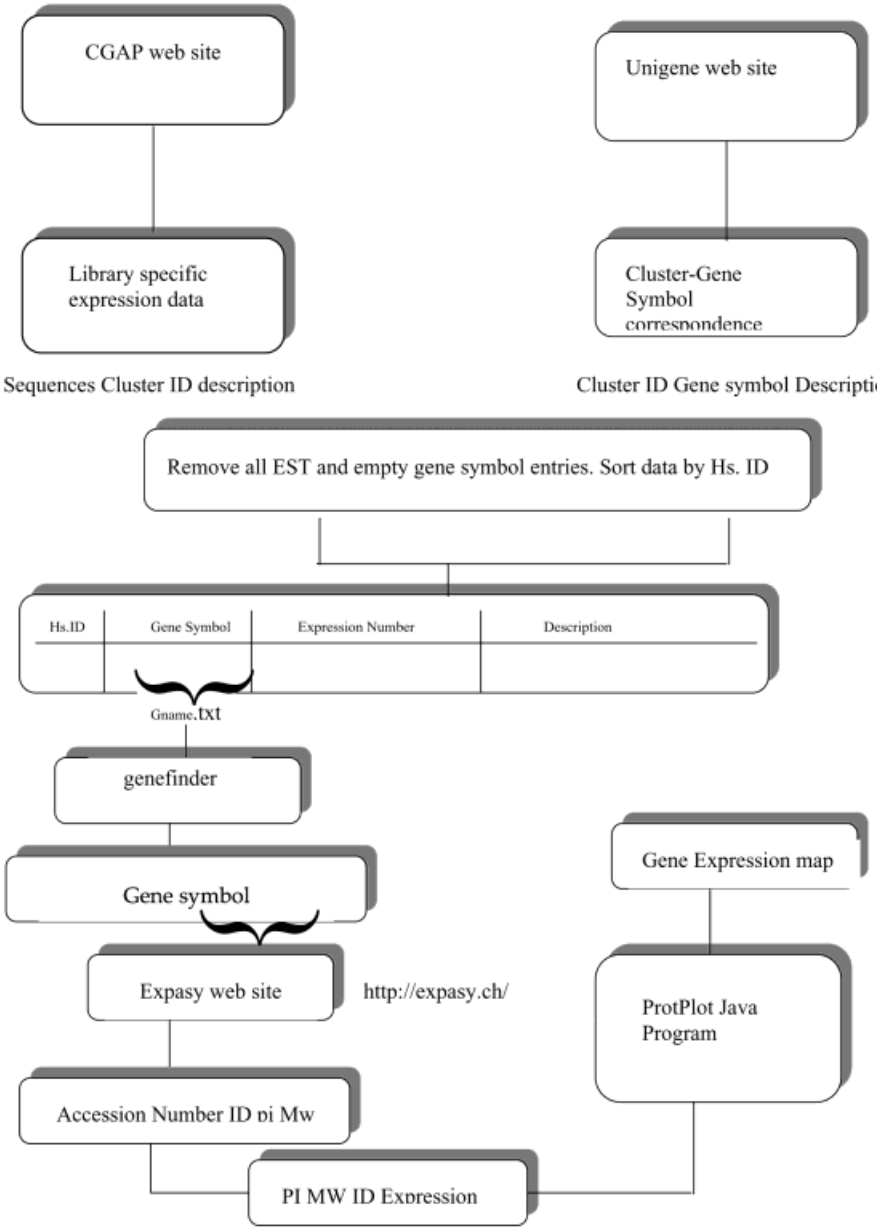

Figure 1. Data mining strategy used to build TMAP expression charts. Tissue specific libraries are first identified. Cross-referencing with SWISS-PROT/TrEMBL using the gene symbol field allows the identification of accession number for each putative, unmodified gene product. This serves as the input to the p*I*/M_r tool server.

option "ANY" from within all 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 interested in computing protein maps, the gene symbols associated with those EST's that were clustered to a gene of known function were extracted from UniGene [5]. 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) [6]. The list of resulting AC was input to the p*I*/M_r tool server which computed the necessary p*I* (isoelectric focusing point) and molecular mass (M_r) for the mature, unmodified proteins [1]. The flow chart is depicted in Fig. 1.

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

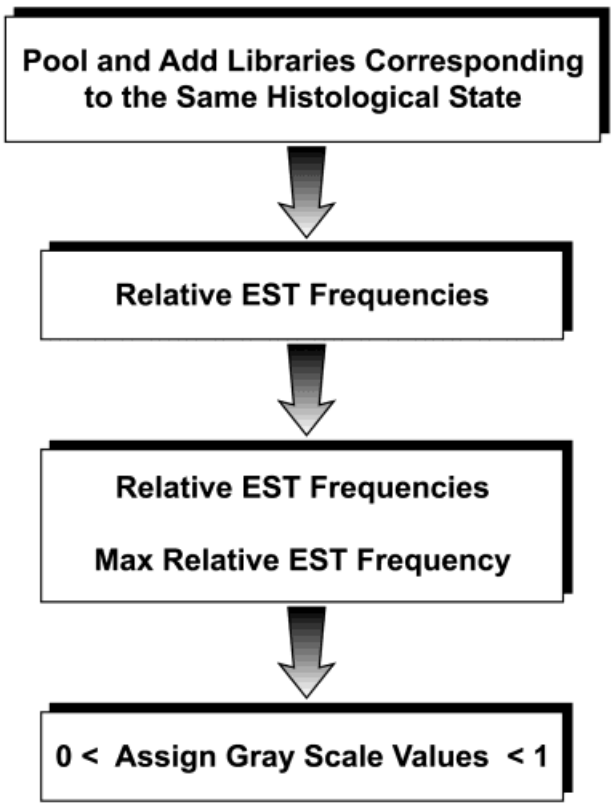

Figure 2. Dataflow to compute relative expression levels from CGAP libraries. Nonredundant lists are computed for pooled datasets.

levels by the maximum relative expression level. In the event a tissue search revealed several libraries fulfilling the requirements of the initial query, to improve the signal-to-noise ratio (S/N), the results are first pooled so as to generate a nonredundant list of entries and a more comprehensive expression map for that tissue and corresponding to that histological state (Fig. 2).

The final result is a tab-delimited file 'prp' format that contains expression levels ranging from 0 (undetected) to 1 (most abundant). The data that is read by ProtPlot when launched loads the set of prp files, creates a master protein index and provides a graphical user interface to perform data mining. ProtPlot is a Java program written by Peter Lemkin and is based largely on Java code from the open-source MAExplorer [7] (maexplorer.sourceforge.net).

The protein data for each tissue is used in constructing the master protein index where proteins will be present for some tissues and not for others. Data is presented as a pseudo 2-D gel image with the isoelectric point (p*I*) on the horizontal axis and the molecular mass (M_r) on the vertical axis. Sliders on each of these axes allow zooming a subregion of interest by restricting the data between the minimum and maximum values of the attributes displayed. By clicking on any of the hyperlinked spots in the scatter plot, information relevant to that protein is displayed. If the genomic-ID web browser is enabled and one is connected to the internet, a web page from the selected genomic database for that protein will pop up.

Samples can be viewed in the pseudo 2-D gel using one of several options. The tissue and histology can be selected using the (File menu | Select Current PRP sample). Two types of scatter plot displays are available: expression data for a single sample (current PRP sample) or the ratio of an X and Y sample (*i.e.*, exprX /exprY). The X (Y) data is selected using (File menu | Select X (Y) PRP sample). An alternative display called the Expression Profile (EP) is a list of a subset of PRP samples. The EP samples can be grouped by histological state, the choice can be specified using the (File menu | Select Expression List of samples) command.

The data is passed through a data filter consisting of the intersection of several tests including: p*I* range, M_r range, sample expression range, expression ratio (X/Y) range (either inside or outside the range), tissue type filter, protein family filter, and clustering. Most of these features are currently available while a few clustering and family classification methods will be implemented in future versions.

At the end of a session, all current settings of data mining strategies can be saved for future use. In addition, to enable information exchange in collaborative efforts, ProtPlot can generate tab-delimited reports as well as pop-up (expression profile) plots and scatter plot (as .gif image files) of proteins satisfying specific search criteria. These can be saved into the project's /Report directory.

3 Results

3.1 Statistical analysis: Prostate, a case study

Prostate is one of the better represented tissues in the CGAP database, offering several libraries spanning normal, precancer as well as cancer states. The availability of these multiple, independent data sets drastically enhances overall quality of the statistical analysis by improving the S/N. Therefore it was selected as the "testing" ground for a number of approaches used to analyze the data.

Thirteen CGAP microdissected prostate libraries encompassing normal, precancer and cancer were found (Table 1). After being grouped according to their histological states, nonredundant lists of the gene-products found within each one was used to compute the p*I*/M_r charts.

Table 1. Microdissected prostate CGAP libraries used to construct an overall p*I*/M_r map for each histopathological state. Normal (N1–N4), precancer (P5–P9) and cancer (C10–C13). The last column represents the list of gene products used to construct each chart.

Dataset	UniGene ID	CGAP Library ID	Patient	Sample type	Gene products
N1	281	NCI_CGAP_Pr1	1	Normal epithelium	3034
N2	515	NCI_CGAP_Pr5	2	Normal epithelium	940
N3	526	NCI_CGAP_Pr9	3	Normal epithelium	1218
N4	529	NCI_CGAP_Pr11	4	Normal epithelium	1354
P5	282	NCI_CGAP_Pr2	1	Premalignant lesion	3303
P6	511	NCI_CGAP_Pr6	2	Premalignant lesion	1343
P7	538	NCI_CGAP_Pr7	2	Premalignant lesion	723
P8	544	NCI_CGAP_Pr4	1	Premalignant lesion	834
P9	545	NCI_CGAP_Pr4.1	1	Premalignant lesion	1132
C10	283	NCI_CGAP_Pr3	1	Adenocarcinoma	2442
C11	513	NCI_CGAP_Pr8	2	Adenocarcinoma	625
C12	523	NCI_CGAP_Pr12	5	Adenocarcinoma	2362
C13	527	NCI_CGAP_Pr10	3	Metastatic adenocarcinoma	747

In addition, using ProtPlot one can compute the ratio of any two maps, which will obviously be carried out over those entries found in both charts. Finally interactive expression profiles across all preselected samples (in this case, cancer only) can be displayed on the fly. Using these tools, one can readily flag gene products with a significant differential expression for postanalysis experimental verification.

One such example is the 60S ribosomal protein L41 (SWISS-PROT ID RL41-HUMAN) whose relative expression appears to steadily increase from the normal to the cancerous state as shown Fig. 3. Additional information regarding this protein, obtained by clicking the corresponding spot which is hyper-linked to a database of choice (*e.g.*, SWISS-PROT) reveals it to be a modulator of cyclin-dependent kinase II, one of the major elements of cell cycle and cell proliferation [10].

When multiple libraries are pooled, the basic assumption is that the individual libraries represent unbiased samples from a "true" dataset that contains all of the expression levels for a particular tissue in a given histological state. This basic assumption can be tested using the Fisher Exact Test to determine the probability that the expression level of a given gene is the same in two datasets. For any given pair of pool sizes (N, M) and gene counts (c and C) the probability p of the table being generated by chance is calculated (see table 2) where:

$$p = \frac{[\mathrm{N!\ M!\ c!\ C!}]}{[(\mathrm{N}+\mathrm{M})!\ \mathrm{a!\ b!\ A!\ B!}]} \quad (1)$$

The null hypothesis of a gene being equally represented in two pools is rejected when probability $p <= 0.05$, where 0.05 is the level of statistical certainty (95% confidence level). Furthermore if Ψ denotes the number of observations of a specific gene in a given library, where the total number of observations of all genes is represented by n, the probability of detecting that particular gene product is simply the ratio:

$$\pi = \frac{\Psi}{n} \quad (2)$$

An obvious normalization condition is:

$$\Sigma\,\pi = 1 \quad (3)$$

An estimate of the uncertainty in this probability at the 95% confidence level is given by [9]

$$\triangle\pi = 1.96\,[\pi - \pi^2/n]^{1/2} \quad (4)$$

This can be used to place 95% confidence limits on the number of observations of a specific gene.

$$\text{Probability of observation} = \pi \pm 1.96\,[\pi - \pi^2/n]^{1/2} \quad (5)$$

$$= \{\Psi \pm 1.96\,[\Psi\,(1 - \Psi/n)]^{1/2}\}/n$$

For small numbers of observations, Ψ/n will be much smaller than one and can be removed from the equation above. This produces:

$$\text{Number of observations} \sim \Psi \pm 1.96\,[\Psi]^{1/2} \quad (6)$$

The following table depicts how this translates for low-abundance gene products.

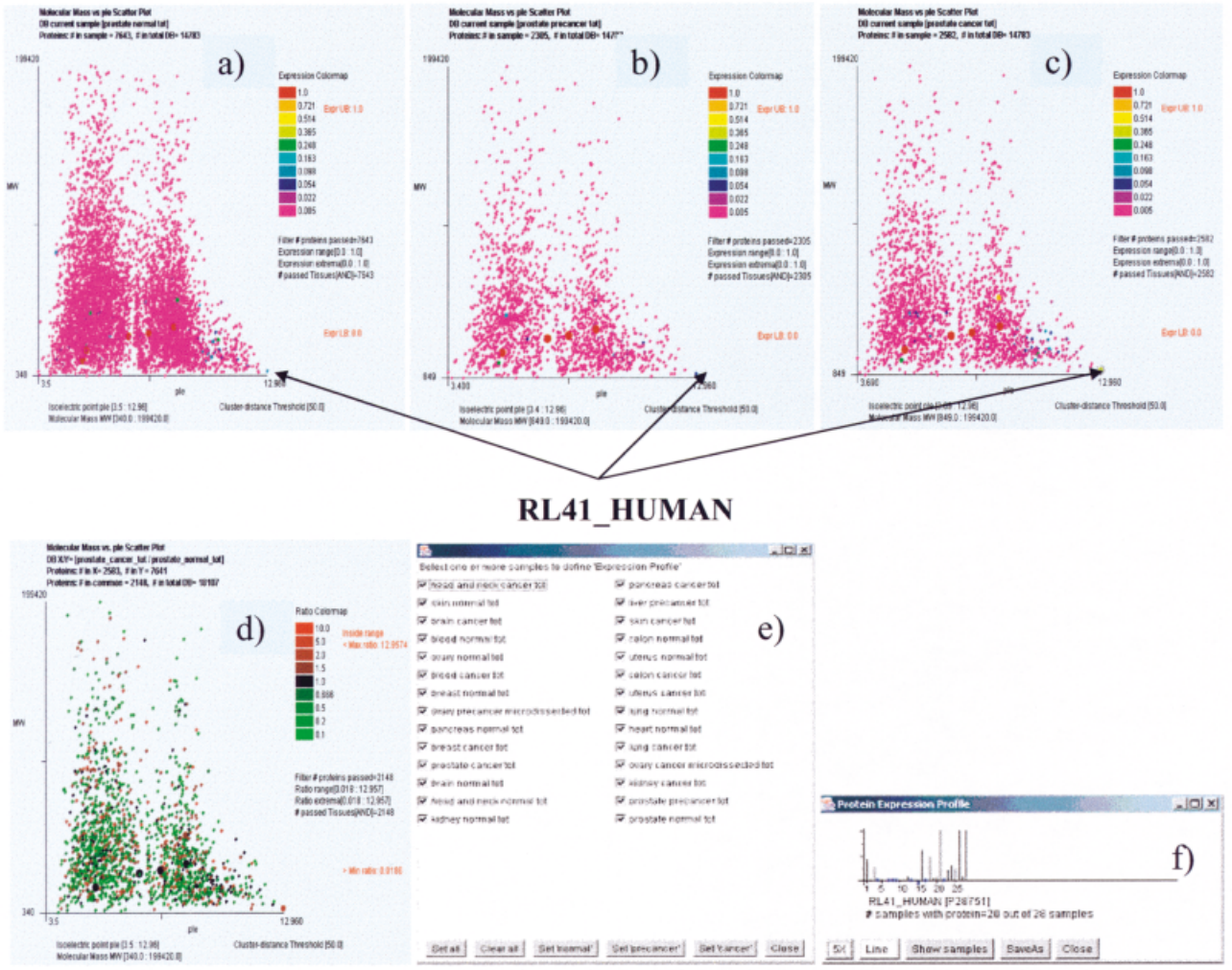

Figure 3. TMAP screenshots for prostate through (a) normal; (b) precancer; and (c) cancer histological states. The respective ranges for the isoelectric focusing point (p*I*) on the horizontal axis and molecular weight (M_r) on the vertical axis correspond to the boundaries of the dataset. The relative expressions of the gene products computed as described in the text are grouped and color-coded as follows: 1.0 0.721 0.514 0.365 0.248 0.163 0.098 0.054; (d) represents the ratio cancer over normal for those gene products detected in both states; (e) Through this panel the user is able to select both tissue and state (here, set for all); (f) expression profiles across all set tissues.

Ψ	△Ψ
1	1.96
2	2.77
3	3.39
4	3.92
5	4.38

This clearly demonstrates that for any gene product, if the number of times it has been detected is less than 5 (in an appropriately sized pool of transcripts), the uncertainty in that measurement is of the same magnitude as the measurement. This can be summarized as follows: low abundance (detected less than five times) or even undetected are approximately equally unreliable and therefore ought not to be included in any statistical analysis.

3.2 Cluster analysis

Another way to obtain this result is to use the Fisher formula and determine the probability that "a" observations of a gene is statistically (95%) the same as not observing it at all, in the same dataset. Therefore, if b = 0, B = M = N, c = a, C = 2N – a, the preceding formula for the probability of "a" and "0" representing the same result is

$$p = (N!/(N-a)!)\ ((2N-a)!/(2N)!) \qquad (7)$$

Table 2. Definition [4] of the terms used in equations (1–6). Given two pools (A, B) containing respectively N, M sequences in total.

Pool A	a = Number of sequences in pool A assigned to Gene 1	A = Number of sequences in Pool A NOT assigned to Gene 1	N = a + A
Pool B	b = Number of sequences in pool B assigned to Gene 1	B = Number of sequences in Pool B NOT assigned to Gene 1	M = b + B
Total	c = a + b	C = A + B	N + M = c + C

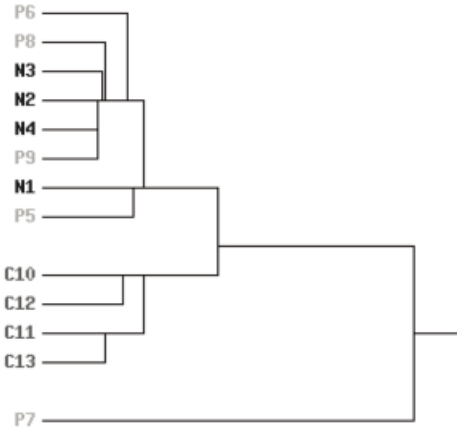

Figure 4. Clustering dendogram of microdissected prostate libraries according to the average *p*-value among gene products detected five or more times.

which reduces to

$$p = [(N)(N-1)\ldots(N-a+1)]/[(2N)(2N-1)\ldots(2N-a+1)] \quad (8)$$

For reasonably large N, this result is very close to

$$p = (1/2)\text{^}a \quad (9)$$

If *p* has to be less than 0.05, then "a" must be 5 or more.

This "minimum of five observations" rule can be used in the comparison of two datasets. For example, the compiled results for prostate generated 13 datasets; four representing gene expression in normal cells (N1 through N4), five representing gene expression in precancer cells (P5 through P9), and four representing gene expression for cancer cells (C10 through C13). The Fisher Exact Test can be used to determine whether these 13 datasets are sampled from the same or different "true" expression datasets.

In particular, two datasets are compared on a gene-by-gene basis. If the observation level is less than 5 for a particular gene in both datasets, that gene is ignored since a statistically significant difference does not exist. Otherwise, the Fisher *p*-value is calculated. The average of these *p*-values represents the average probability that these libraries are statistically similar for the set of genes

Table 3. Pairwise comparison of the expression for each gene product expressed more than five times in any dataset for all libraries. (Set 1 denotes Normal 1 (N1), set 2 denotes Normal 2 (N2), etc.).

Set	Set	AVG(P)	Number
1	2.3	3385	116
1	3.2	3867	116
1	4.1	8101	121
1	5.0	7700	187
1	6.1	9752	117
1	7.5	3119	112
1	8.4	0839	113
1	9.2	6353	114
1	10.0	1336	231
1	11.1	9340	130
1	12.0	1357	207
1	13.1	5596	139
2	3.0	9598	19
2	4.1	0614	27
2	5.3	7594	137
2	6.0	6117	17
2	7.1	7255	9
2	8.0	7436	11
2	9.0	9380	16
2	10.3	3458	128
2	11.0	3681	27
2	12.2	6915	104
2	13.0	4470	36
3	4.0	9349	28
3	5.2	6347	137
3	6.0	5776	21
3	7.2	2508	14
3	8.1	2551	17
3	9.0	9850	22
3	10.2	0063	133
3	11.0	1553	32
3	12.1	4508	109
3	13.0	1593	41
4	5.2	2017	139
4	6.0	5583	28
4	7.2	6219	19
4	8.1	7250	22
4	9.0	9973	27

Table 4. Interlibrary similarity can be evaluated by computing the distance matrix displaying the average *p*-values of all thirteen prostate datasets when compiled over all gene products expressed five or more times in each library.

	N1	N2	N3	N4	P5	P6	P7	P8	P9	C10	C11	C12	C13
N1	1.00000	.09262	.06673	.07534	.10599	.08484	.00000	.00658	.06775	.00000	.00000	.00000	.00000
N2	.09262	1.00000	.16833	.17918	.05304	.01647	.00000	.08186	.10440	.00000	.00000	.00000	.00000
N3	.06673	.16833	1.00000	.16959	.06946	.07344	.00000	.00000	.10135	.00000	.00000	.00000	.00000
N4	.07534	.17918	.16959	1.00000	.05743	.11575	.00000	.00000	.18116	.00000	.00000	.00000	.00000
P5	.10599	.05304	.06946	.05743	1.00000	.06729	.00000	.00188	.06813	.00000	.00000	.00000	.04098
P6	.08484	.01647	.07344	.11575	.06729	1.00000	.00000	.00000	.00070	.00000	.00000	.00000	.00000
P7	.00000	.00000	.00000	.00000	.00000	.00000	1.00000	.00000	.00000	.00000	.00000	.00000	.00000
P8	.00658	.08186	.00000	.00000	.00188	.00000	.00000	1.00000	.15987	.00000	.00000	.00000	.00000
P9	.06775	.10440	.10135	.18116	.06813	.00070	.00000	.15987	1.00000	.00000	.00000	.00000	.00000
C10	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	1.00000	.09014	.12158	.07215
C11	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.09014	1.00000	.09398	.16075
C12	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.12158	.09398	1.00000	.06478
C13	.00000	.00000	.00000	.00000	.04098	.00000	.00000	.00000	.00000	.07215	.16075	.06478	1.00000

that have five or more observations in at least one of the two datasets. These average *p*-values can be used to determine a pseudo-distance between the 13 datasets, which can then be used in a clustering algorithm. This pseudo-distance, D, between two datasets is obtained from their average *p*-value, ρ, using the expression

$$D = [(1 + \delta)/(\rho + \delta)] - 1 \quad (10)$$

In this expression δ is a constant so that when $\rho = 0$, $D = 1/\delta$. With values of D for all pairs of datasets, an agglomerative clustering can be performed. A dendogram showing the results from single linkage clustering of the 13 datasets is presented in Fig. 4. This dendogram uses $\delta = 0.04$ in the above distance equation, so the maximum separation between datasets that have no chance of representing the same results is 25.0.

Table 3 compares all pairs of datasets to determine the extent to which they are statistically similar. This is done by looking at all genes that are expressed in the two datasets (five or more observations) and by determining the average *p*-value for this set of expressed genes. The list at the top of the file gives the number of the first dataset, the number of the second dataset, the average *p*-value for the genes that are expressed in each set, and the number of genes used for this average. In Table 4 we compute the average *p*-value matrix for these 13 datasets. These results show that the normal datasets (first four) agree with each other with an average probability of 6.67% or higher. Similarly, the cancer datasets (last four) agree with each other with an average probability of 6.48% or higher. There is a 0.0 probability that a normal dataset and a cancer dataset are samples of the same, large dataset. In other words, these samples came from different pools. A closer examination of the top part of the output shows that no gene is significantly expressed in both a normal (sets 1–4) and a cancer (sets 10–13) dataset. Every precancer dataset (5–9) but #7 has a significant agreement with at least one normal dataset, and no agreement with any cancer dataset. Dataset 7 either has no significantly expressed genes, or has no significantly expressed gene that is not significantly expressed in any other dataset (the union of expressed genes from 7 and any other dataset is the empty set).

4 Discussion

Figure 5 depicts the p*I*//M_r 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 alphabet [1]. 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.

Finally, given a reduced number of histological states, the number of possible scenarios for the behavior of expression of any gene product is finite. If normal and cancer states are the only options, then there are only three possible cases in going from one to the other: (1) remain at the same level, (2) up-regulated; (3) down-regulated. If there are three histological states to consider (normal, precancer and cancer), the number of possible scenarios jumps to nine as can be attested in Fig. 6. One possible application of this approach is that one is able to cluster those gene products that are potentially inversely correlated. One such example is provided by p27, down-regulated in prostate cancer as a result of proteolitic cleavage by an overexpressed enzyme, Skp2.

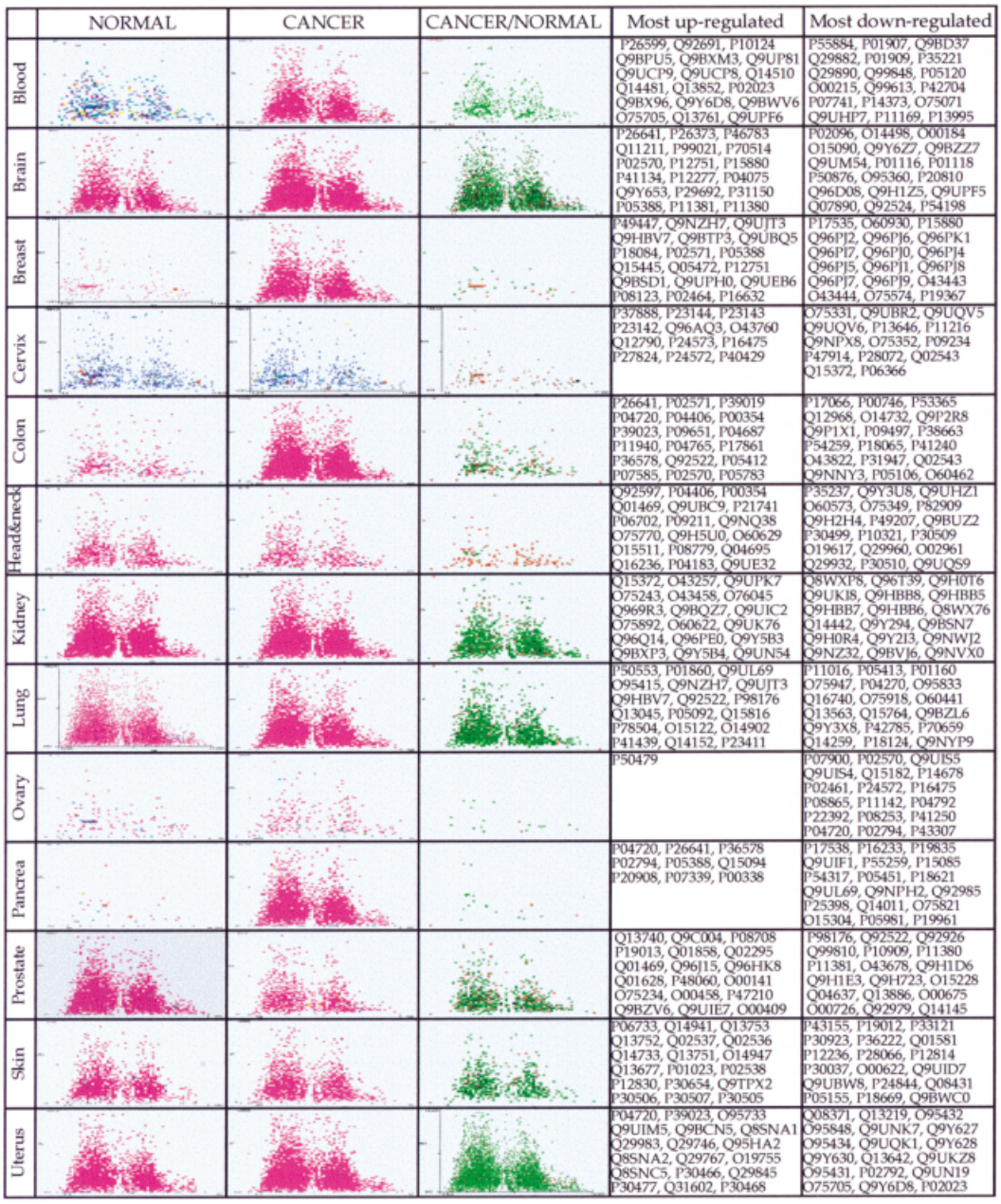

Figure 5. Tissue and histology specific p*I*/*M*$_r$ maps surveyed to date. The color code for scatter plots is the same as in Fig. 4 for the individual maps, but for ratios (X/Y) it is as follows: 10.0 5.0 2.0 1.5 1.0 0.666 0.5 0.2 0.1. The last two columns list the SWISS-PROT accession numbers for those gene products with the highest and lowest cancer/normal expression ratios, respectively.

5 Concluding remarks

In this survey 144 libraries from the CGAP public resource were used to produce more than 18 000 putative gene products encompassing normal, cancerous and, when available, precancerous states for fourteen tissues. It is the aim of this study to lay the foundation of a future extensive analysis of existing public expression databases. Present and future implementations will allow the user to cluster proteins by the similarity of their expression profiles. The clustering method and distance metric based on expression profile are selectable from a number of options in the menu. These interactive, web-able maps and associated graphics Java-based software PROT-

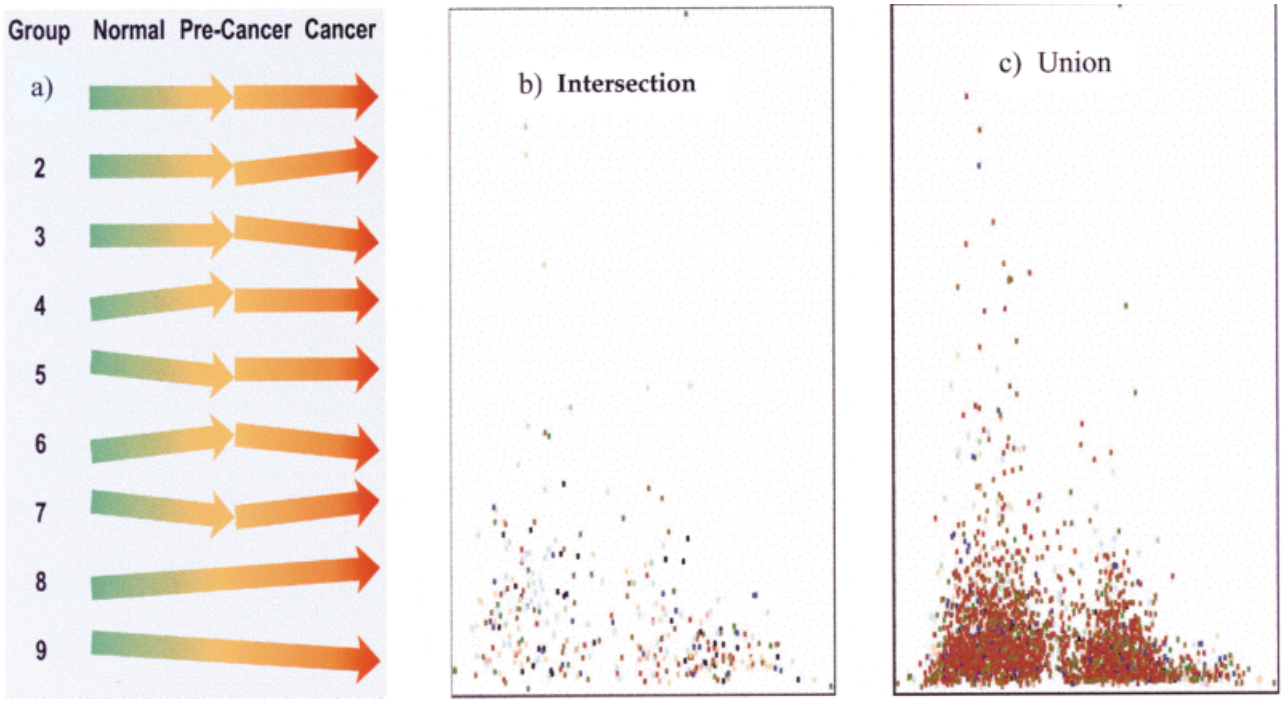

Figure 6. (a) Color-coded symbolic representation of all possible scenarios of expression profile through three histological states N = P = C, N = P < C, N = P > C, N < P < C, N > P > C, N < P = C, N > P = C, N < P > C, N > P < C; (b) color-coded clustering for entries found in all three histological states (intersection); (c) color clustering of non-redundant list found in any of the three histological states.

PLOT collectively known as VIRTUAL2D [1] and TMAP are available to the nonprofit research community upon request.

Received January 10, 2003

6 References

[1] Medjahed, D., Smythers, G. W., Stephens, R. M., Powell, A. D. *et al.*, *Proteomics* 2003, *3*, 129–138.

[2] Grouse, L. H., Munson, P. J., Nelson, P. S., *Urology* 2001, *57*, (*Suppl. 1*), 154–159.

[3] Ryu, B., Jones, J., Hollingsworth, M. A., Hruban, R. H., Kern, S. E., *Cancer Res.* 2001, *61*, 1833–1838.

[4] Cancer genome anatomy project at http://CGAP.nci.nih.gov/

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

[6] SWISS-PROT at http://www.expasy.ch

[7] Lemkin, P. F., Thornwall, G. C., Walton, K. D., Hennighausen, L., *Nucleic Acid Res.* 2000, *28*, 4452–4459.

[8] Lee, J.-H., Kim, J.-M., Kim, M.-S., Lee, Y.-T. *et al.*, *Biochem. Biophys. Res. Commun.* 1997, *238*, 462–467.

[9] Gillespie, J. H. in: *Population Genetics: A Concise Guide,* The Johns Hopkins University Press, Baltimore 1998 p. 8.

[10] http://us.expasy.org/cgi-bin/niceprot.p1?P28751