Classification of Proteins and Biomarkers to Tissue Types Using Publicly Available Gene Chip Data

Poster Number MPB 064

Kenneth C. Parker BG- Medicine, Waltham, Massachusetts USA (now at Virgin Instruments Corporation Sudbury, MA)

Overview

- Assign proteins to the cell type in which they are most abundant
- Download gene chip data for relevant tissues
- Perform PCA and K means clustering
- Annotate results

Introduction

Many proteomics applications have as their goal the identification of protein biomarkers that can be used to monitor disease status, starting from plasma or tissue biopsies. In either case, the samples often consist of proteins originally expressed by many cell types, which may be difficult to separate from one another. The credibility of a biomarker is strengthened if a case can be made that the biomarker is a member of a class of proteins derived from a specific cell type in the diseased tissue. A method is described by which publicly available gene chip data can be used to classify proteins detected by proteomics according to cells, tissues, and organs that are relevant to the disease in question.

Methods

- Example heterogeneous tissue: atherosclerotic plaque
    - Known constituents:
        - endothelial cells
        - smooth muscle
        - infiltrating lymphocytes
        - plasma
- Choose datasets from GEO web site
    - Check proteomics data to ensure that the gene profile datasets express the most abundant proteins
    - Novartis dataset (Su et al.)
    - White blood cell study (Jeffrey et al.)
- Data mining
    - Align proteins using gene symbol, saving the largest expression value per tissue
    - Normalize by column (% expression attributed to each gene for tissue type)
    - Prepare a table of gene vs. % expression
    - Perform K means clustering (60 clusters) and PCA (12 components)
    - Color PCA plots based on K means cluster
- Annotate master gene table according to K means cluster and PCA coefficients

Caveats

- Using proteomics data to accomplish this would be superior
- Such data needs to be shotgun, and retain abundance statistics
    - Couldn't find any
- Gene chip data on more relevant tissues would have been preferable
    - No decent aorta endothelial dataset
    - Some gene datasets don't merge well with others (apples vs. oranges)
-

Table 1. Tissue Gene Profiles

tissue	ref	order	K
adipocytes	Su et al.	16	58
B cells	Jeffrey et al.	15	59
bone marrow cd33 myeloid	Su et al.	14	22
bone marrow cd34	Su et al.	13	46
endothelial_cd105	Su et al.	12	7
erythroid early cd71	Su et al.	11	48
liver	Su et al.	9	60
liver_fetal	Su et al.	10	57
macrophage	Jeffrey et al.	8	52
macrophage stimulated	Jeffrey et al.	7	55
monocytes cd14	Su et al.	6	33
neutrophil	Jeffrey et al.	5	47
neutrophil stimulated	Jeffrey et al.	4	51
smooth muscle	Su et al.	3	56
TH1 cells	Jeffrey et al.	2	54
TH2 cells	Jeffrey et al.	1	50

order: as in Fig 2A-C. **K**: most specific K cluster in other figures and tables

Fig. 1. Histogram of the expression profile of a gene (PRTN3) in a 79 tissue dataset (Su et al.) using the GEO site at NCBI.

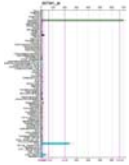

Fig. 2A Expanded view of tissue profile for cluster K1

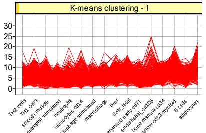

Table 2.
K clusters 41-60 (most tissue-specific);
16 tissues in columns;
K clusters in rows.
Compare deduced profile name to pattern at bottom of Fig, 2B.

Color key: >50%, >30%, >20%, >10% of intensity in indicated tissue.

K	#	type	MF	MF_s	Neu_s	Neu	TH1	TH2	B	smooth	cd105	liver	liver_fetal	adipo	mono	cd33	erythro	cd34
60	71	liver	1.22	1.06	1.23	0.78	1.04	0.97	0.87	1.01	0.61	**73.88**	**10.97**	3.17	0.71	1.01	0.61	0.87
59	33	B	2.01	1.85	1.23	1.02	2.08	2.76	**73.13**	1.00	1.45	2.30	2.48	2.22	1.08	1.42	1.10	2.87
58	60	adipo	2.26	1.75	1.20	0.93	1.84	1.47	1.30	6.15	1.20	3.84	3.23	**71.32**	0.67	1.11	0.70	1.04
57	29	liver_f	1.30	1.37	0.85	0.81	1.20	0.95	1.02	1.07	0.71	**12.24**	**70.64**	3.17	1.04	1.25	1.15	1.23
56	26	smooth	2.18	2.94	1.38	1.01	1.11	1.04	0.82	**69.51**	1.41	1.92	3.89	9.59	0.55	0.89	0.57	1.19
55	55	MF_stim	8.54	**64.30**	3.14	2.69	4.77	1.80	2.05	2.40	0.68	1.23	1.49	2.39	0.91	1.61	0.80	1.20
54	27	TH1_hi	1.26	2.45	1.30	1.01	**60.67**	**17.92**	3.51	1.01	0.86	1.77	2.43	2.17	0.88	0.88	0.65	1.21
53	40	smooth_adipo	1.87	1.65	1.19	0.71	1.32	0.92	0.99	**39.24**	0.97	2.60	4.25	**41.28**	0.58	0.76	0.58	1.08
52	94	MF	**36.87**	**42.55**	1.23	0.98	1.79	1.47	1.65	1.74	0.74	1.69	1.83	2.67	1.34	1.69	0.68	1.07
51	70	Neu_stim	3.75	6.05	**53.22**	**12.25**	3.85	3.10	3.00	1.38	0.64	1.45	1.60	2.05	2.37	3.17	0.74	1.38
50	26	TH2	2.96	2.26	1.67	1.07	**22.77**	**49.54**	2.52	1.31	1.10	2.02	3.34	3.01	0.99	1.40	1.15	2.88
49	60	liver_b	1.93	1.80	1.45	1.02	1.74	1.87	1.26	1.80	1.11	**38.45**	**38.84**	4.27	0.85	1.18	1.27	1.17
48	35	erythro	1.75	1.55	1.16	1.08	2.09	1.61	1.30	1.58	**12.15**	1.77	**16.88**	2.80	1.08	1.22	**48.67**	3.31
47	69	Neu	3.20	2.88	**36.08**	**35.87**	2.64	2.04	2.77	0.92	0.64	1.61	1.48	2.62	2.44	3.20	0.51	1.09
46	19	CD34	1.86	1.98	2.49	2.68	3.78	2.75	3.33	1.75	6.24	2.90	4.92	3.22	2.79	9.13	1.40	**48.77**
45	4	liver_neu	0.94	0.87	**32.80**	**15.75**	1.22	0.71	0.78	2.33	0.57	**34.47**	4.35	2.10	0.85	0.81	0.68	0.78
44	80	liver	4.66	3.86	2.47	1.87	3.23	3.33	2.56	2.65	1.81	**46.36**	**11.18**	7.51	2.08	2.44	1.52	2.47
43	112	MF_un	**41.97**	**23.72**	1.89	1.69	3.47	3.47	2.54	2.31	1.07	2.50	3.01	4.68	2.02	2.60	0.93	2.14
42	47	MF_stim_neu_stim	9.50	**32.70**	**27.87**	**11.56**	4.05	2.59	2.15	2.04	0.37	0.64	0.81	1.27	1.47	2.05	0.24	0.67
41	156	T	5.73	3.70	1.57	0.95	**32.88**	**29.23**	3.57	1.70	3.10	1.37	3.57	2.10	0.94	1.77	3.04	4.78

Fig. 2A Expanded view of tissue profile for cluster K60

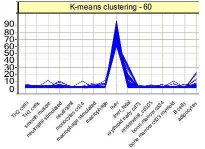

Table 3.
Deduced tissues that dominate each of the 60 K clusters.
# indicates the number of genes in the cluster.

K	#	type
1	391	common
2	595	MF_T_B_common
3	399	Neu_stim_common
4	879	liver_f_adipo_comm
5	333	MF_not_cd105_com
6	355	T_B_common
7	192	liver_f_adipo_comm
8	372	T_common
9	223	B_common
10	283	blood_common
11	414	blood
12	459	liver_f
13	433	adipo_liver
14	89	MF_cd33_common
15	74	erythro_liver_f
16	636	T_MF_B
17	212	MF_common
18	153	endo5
19	294	liver_b_adipo
20	188	B
21	385	B
22	25	CD33
23	340	MF_unstim_T
24	160	MF_neu
25	66	MF_smooth
26	135	TH1
27	206	MF_B
28	385	T
29	342	MF_T
30	173	Neu_stim
31	126	Neu_unstim
32	155	MF_stim_lo
33	23	mono
34	51	Neu_both
35	246	MF_both
36	70	smooth
37	72	liver_f
38	123	B
39	147	adipo
40	7	CD105
41	156	T
42	47	MF_stim_neu_stim
43	112	MF_un
44	80	liver
45	4	liver_neu
46	19	CD34
47	69	Neu
48	35	erythro
49	60	liver_b
50	26	TH2
51	70	Neu_stim
52	94	MF
53	40	smooth_adipo
54	27	TH1_hi
55	55	MF_stim
56	26	smooth
57	29	liver_f
58	60	adipo
59	33	B
60	71	liver

Fig. 2B.
Tissue profiles for each of the 60 K clusters.

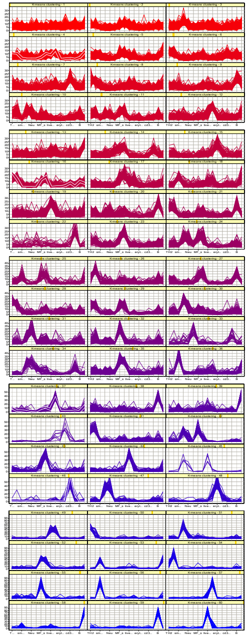

Fig. 3.
Histogram of gene occupancy of K clusters.

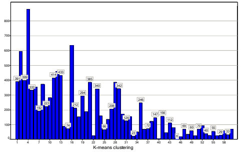

Fig. 4A-B.
PCA plots colored by K cluster. Each pole tends to be enriched in a particular tissue. Other tissues dominate (as listed in Table 4) when other PCA dimensions are plotted.

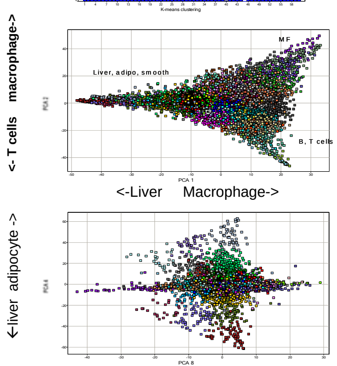

Fig. 4C.
Exploded view:
PCA4 vs. PCA8, 60 K clusters, 10 left on extremes
639 remaining of 11324 genes.

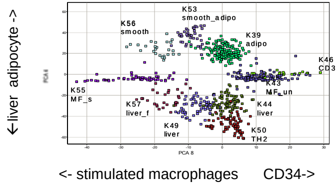

Table 4.
Annotated Scree plot showing PCA dimension, and relationship between tissues and PCA dimension.

Scree Plot (Eigenvalues)					
Principal Component	Eigenvalue	Eigenvalue (%)	Cumulative Eigenvalue (%)	hi	lo
PC (1)	163.785	25.4	25.4	MF	liver
PC (2)	101.818	15.8	41.2	MF	T
PC (3)	85.870	13.3	54.5	MF	neutrophil
PC (4)	68.733	10.7	65.1	adipocyte	liver
PC (5)	52.557	8.1	73.3	T	B
PC (6)	37.328	5.787	79.077	liver	fetal liver
PC (7)	27.99	4.339	83.416	adipocyte	smooth
PC (8)	26.54	4.114	87.531	CD34	MF_stim
PC (9)	23.279	3.609	91.14	smooth	MF; CD34
PC (10)	14.7	2.279	93.418	neu_stim	neu_unstim
PC (11)	13.598	2.108	95.526	erythro	CD34; liver;mono
PC (12)	11.684	1.811	97.338	TH2	TH1

Conclusions

- Can assign many genes (proteins) to specific tissues
- Most housekeeping proteins have a particular uneven distribution
- Each PCA dimension maps to a particular combination of tissues
    - (~ as expected for principal component analysis)
- Each K cluster consists of proteins with a similar tissue distribution
- This same process works based solely on proteomics data
- The input datasets need not be pure
    - The process works even when input data consists of replicates that have unintended residual variability

References

- Andersen, J. S.; Wilkinson, C. J.; Mayor, T.; Mortensen, P.; Nigg, E. A.; Mann, M., (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 426, 570-4.
- Jeffrey KL, Brummer T, Rolph MS, Liu SM, Callejas NA, Grumont RJ, Gillieron C, Mackay F, Grey S, Camps M, Rommel C, Gerondakis SD, Mackay CR (2006) Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1., Nat Immunol. 7:274-83.
- Parker KC, Walsh R, Salajegheh M, Amato A, Krastins B, Sarracino D, Greenberg SA (2009) Characterization of human skeletal muscle biopsy samples using shotgun proteomics. J Proteome Res. Apr 21. (somewhat similar methodology applied to muscle biopsy)
- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 101:6062-7.

Acknowledgments

- Spotfire for graphics and statistical software
- Microsoft for Access and Excel
- NCBI for the GEO datasets
- BG-Medicine for letting me study protein informatics
- Virgin Instruments for paying my way here