LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-226770

Pathomics: Final Report

K.W. Turteltaub, M. Ascher, R. Langlois, I. Fodor, J. Kercher, K. Mc Laughlin, D. Nelson, W. Colston, F.P. Milanovich

December 13, 2006

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Pathomics
Final Report
12/3/2006

Kenneth Turteltaub, Michael Ascher, Richard Langlois, Imola Fodor, James Kercher, Kevin McLaughlin, David Nelson, Bill Colston, and Fred Milanovich

Purpose

Pathomics is a research project to explore the feasibility for developing biosignatures for early infectious disease detection in humans, particularly those that represent a threat from bioterrorism. Our goal is to use a science-based approach to better understand the underlying molecular basis of disease and to find sensitive, robust, and specific combinations of biological molecules (biosignatures) in the host that will indicate the presence of developing infection prior to overt symptoms (pre-syndromic). The ultimate goal is develop a national surveillance system for monitoring for the release and managing the consequences of a biothreat agent or an emerging disease. Developing the science for a more comprehensive understanding of the molecular basis of infectious disease and the development of biosignature-based diagnostics could help detect both emerging and engineered treats to humans.

Background

It has been well-described that the health and societal consequences of the release of a bioterror agent or the emergence of a new natural pathogen could be severe (see Kaplan et al., 2002; O'Toole et al., 2002 for examples). Whether through an act of terrorism or nature, infectious disease epidemics are potentially the most lethal and certainly most insidious of natural disasters (Nelson et al. 2001). Bubonic plague (*Yersinia pestis*) was responsible for a staggering 25 million deaths (roughly a quarter of the entire population) in 14th century Europe. Although less deadly than the plague, smallpox had a tremendous impact on the development of Western civilization (Barquet et al. 1997; Nelson et al. 2001). In 1918 a global pandemic caused by a particularly virulent strain of influenza killed more than 40 million people in the span of 8 months and hospitalized more individuals than the total number of those wounded in World War I (McConnell 2002; Meltzer et al. 1999).

The development of antibiotic and vaccine therapies in the early 1900's resulted in a significant decrease in mortality from infectious disease. However, in recent years the U.S. death rate from infectious disease has begun to rise again (Armstrong et al. 1999). Influenza and pneumonia remain among the top ten causes of death for all age classes in the United States (Anderson 2001; Snacken et al. 1999) and new emergent infectious diseases has posed serious threats to public health (Binder et al. 2002; Noah et al. 2000).

West Nile virus, for example, which broke out in Romania in 1996 and Russia in 1999, has recently spread throughout most of the 48 Continental United States, with more than 3500 reported human cases and 211 fatalities as of November 2002 (Editors 2002). Perhaps the most devastating infectious disease that humanity has faced since smallpox and bubonic plague, Acquired Immune Deficiency Syndrome (AIDS) has struck 60 million individuals worldwide. Five million new cases of HIV infection were reported in 2001, with 3 million deaths and 40 million individuals living with HIV/AIDS. Furthermore, some existing pathogens are becoming more virulent and less sensitive to existing treatments; and genetic engineering techniques now enable the creation of even more deadly pathogens (Cello et al.,). The use of biological pathogens in warfare has presented additional challenges to the public health systems. Use of crude forms of bioagents are evident in the early 14th century (de Lorenzo and Porter, 2000; Inglesby et alk, 2000) and the events of October, 2001 had demonstrated that the potential for real and extensive public and economic impacts are real from bioterrorism.

Acknowledgment of the threats discussed above by the public health community has generated much discussion on how to prepare including the potential for use of host-based surveillance systems that could provide early warning of an attack or new natural pathogen and help avert massive casualties. A number of potential bioterror agents, including Bacillus anthracis (anthrax), Yersinia pestis (plague), botulism and Variola major (smallpox), can be treated successfully if they are diagnosed early. However, they also progress quickly from mild symptoms to serious illness to death, so a quick diagnosis is vital to limit mortality.

Traditionally, health departments have relied on astute doctors to identify emerging pathogens and bioterror attacks by diagnosis. Dr. Larry M. Bush, a physician at JFK Medical Center in Atlantis, Florida, identified anthrax in Bob Stevens, a photo editor for a supermarket tabloid. However, this approach provides indications of an incident only after individuals become overtly symptomatic, when it may be too late for effective treatment (as was the case for Bob Stevens and several other victims). What is needed is diagnostic methods that can indicate that a person is getting sick and identify the causative agent quickly and at the time the individual reports feeling ill. Very frequently these symptoms are general and non-specific.

One new approach to earlier disease detection is to sample the changes that occur in the host at the molecular level in response to the stress caused by invasion by a pathogen. Stressors can cause underlying changes in the biochemisty of the host that result in increases or decreases in the levels of certain biomolecules and appearance of new biomolecules that could potentially be detected in blood or other tissues. We suggest that the characteristics of this molecular response could be dependent on the pathogenesis of the infection resulting in a unique biosignature based on the mechanism of virulence and pathogenesis. Such specificity in response and early detection strategies have been suggested for cancer and other diseases. Thus we propose here a fundamental change in philosophy. Rather than continue the current approach of detecting threats by identifying them as a specific microbe in a taxonomy (e.g., this bacterium is Bacillus anthracis), we propose to detect them based on their mechanism of pathogenesis using host-based

molecular signatures. To ultimately accomplish this goal, a better understanding of the mechanisms of pathogenicity as well as host responses and susceptibilities is needed at the molecular level. Instruments and protocols must be developed that apply the science to surveillance and other monitoring strategies that have dual-use in national security and public health. This mechanism-based approach to detection and characterization of biothreats also has a natural consequence of identifying biological pathways as targets for therapy and result in new therapeutics.

The purpose of this project then, is to begin to develop an approach to discover host-based biomolecular signatures that are indicative of a developing infectious disease and explore how early in the infectious disease course the signature can be detected. We hypothesize that much of this response is programmed by the genome and thus is an inherent property of the host. Understanding this response could be diagnostic of a pathogen's presence and identity.

This initiative is intended to develop the team and preliminary data to carry out this decade long vision leading ultimately to diagnostics for the biological threats from terrorism and emerging diseases. Here we focus on the immediate national need to demonstrate that molecular diagnosis is possible, provide an indication of how early it is possible and suggest a diagnostic platform that could carry out this mission. The capabilities developed via this initiative also have application to developing the medical countermeasures needed for the chemical and radiological/nuclear threat, potentially as an "all hazard" biodosimetry or medical forensic tool.

Approach

The central theme of this project was to determine whether it is feasible to discover and utilize molecular signatures from blood to differentiate infected from un-infected healthy individuals. It is the projects hypothesis that molecular signatures, if measurable, could detect developing infection in the prodromal or presyndromic period. Detection of a developing infection at these stages would offer savings in morbidity and mortality due to the ability to intervene earlier than now possible.

Key challenges to success were the sensitivity and reproducibility of current technology, natural variation in the molecular components that make up the signatures, the extend of change in the molecular signatures to be measured between health and infected individuals, and the ability to process, sort, analyze and find molecular signatures in the large amount of data to be gathered.

The approach undertaken was a tightly coupled experimental and statistical/informatics plan with 3 specific questions ultimately being addressed:

1. Evaluate and select methods for the high throughput screening of blood for molecular signature characterization (year 1 – 2).

2. Characterize the normal range for candidate host response surrogate markers in humans and animals (year 2 – 3).

3. Differentiate the host response surrogate marker profile of infected sick or presyndromic animals or humans (year 2 -3).

Two additional aims were originally proposed but were not executed do to time limitations. These should be the goals of future projects.

- Differentiate the host response surrogate marker profile to a bacterial infection compared to a viral infection.

- Differentiate the host response surrogate marker profiles to different biothreat organisms.

To address these questions, animal models (mouse) and humans (where appropriate) were sampled during the course of an infection and when healthy to discover those molecules that change in concentration in response to the infectious agent. Samples were to be analyzed to provide a molecular target pool from which to discover signatures of the disease. Candidate signatures were evaluated for sensitivity, specificity, and relevance to human infections. Signatures were to be developed into assays and validated for use in later human studies (Figure 1).

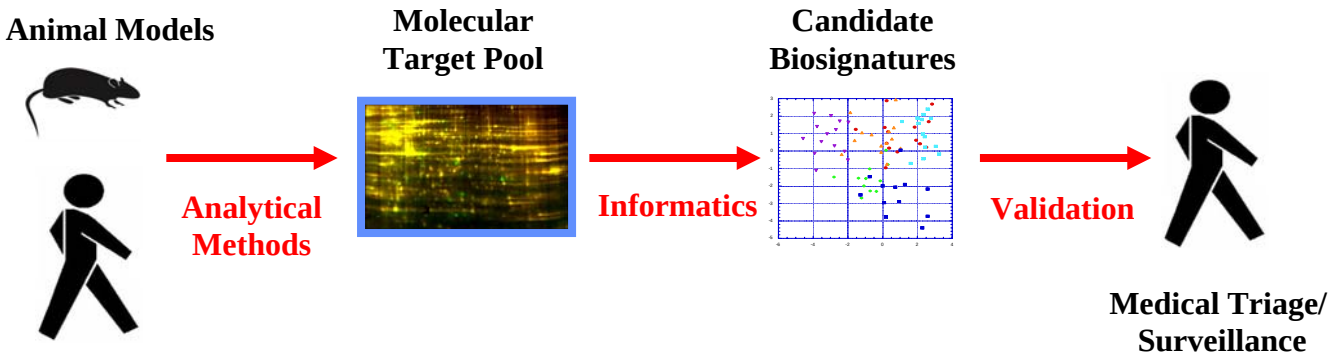


Figure 1: Iterative and coupled experimental and computational approach.

The analytical approach was to use multiplex and/or high throughput technologies such as microarrays (broad screen for gene expression levels), RT-PCR (Real-time polymerase chain reaction), MAP (Multi-Analyte Profiles of proteins), and 2D-DIGE (2-Dimensional Difference In Gel Electrophoresis) to study the molecular response to the pathogen. Unlike the examination of single genes or proteins in isolation as has been done in traditional biology, systems biology simultaneously studies the complex interaction of

many levels of biological information (DNA, mRNA, metabolites and proteins). While the scope of this endeavor is truly a long-term grand challenge, we focused our efforts during this two-year period to develop key analytical, experimental, and computational tools for use with nucleic acid and protein signatures from blood.

First, 2D-Dige, RT-PCR and MAP technologies were evaluated for sensitivity, precision and reproducibility. The central challenge is the ability of these methods to discriminate changes in molecular concentration. These results are presented under Technical Accomplishments - aim 1.

Second, RT-PCR and MAP technologies were determined to be the ideal analytical platforms at this time since these are the most well characterized and are quantitative, providing the ability to not only qualitatively state that a component of a signature is present but also to quantify the change in concentration. This added dimension of quantitation over a binary analysis of presence v. absence may add specificity to the signature. Statistical methods were also evaluated for use in analysis of these data sets. Results are presented under Technical Accomplishments – aim 1).

Third, a study was conducted to define the normal 'healthy' individual on a molecular basis. This is necessary to determine what molecular changes are useful in a signature for disease detection. It is also necessary to understand the normal variation in these signatures among the population and over time to make sure that the analytical parameters of the methods used are capable of detecting the level of changes encountered and to set criteria for what constitutes a useful signature component. These results are presented under Technical Accomplishments – aim 2.

Fourth, we began to assess how early in a disease process these methods could discriminate health from infected individuals. Since a controlled challenge with a threat agent is not possible in humans, an animal model was chosen for feasibility assessment. Using a mouse cowpox model and a multiplex bead-based proteomic assay system which measures over 100 cytokines and chemokines simultaneously, we have shown that host responses are detectable in serum from a localized lung infection 1-2 days prior to overt signs of illness. The pattern of response or signature varied among markers presenting the possibility of using the pattern to determine the stage of infection (see Technical Accomplishments – aim 3). In addition, we collaborated with a dialysis network to collect samples from humans undergoing dialysis. These individuals were ideal for a feasibility study because they are sampled 3-times a week. Samples could be banked for a set of individuals and analyzed once they became sick. Baselines were constructed from periods when they were apparently infection free and a time-course could be constructed from the routine samplings. Approximately 1200 samples were collected and banked over a years time period. Longer term studies were planed in humans with natural opportunistic infections once the methods were demonstrate and validated in an animal model.

List of Technical Accomplishments (see appendix for manuscripts)

Aim 1. Evaluate and select methods for the high throughput screening of blood for molecular signature characterization (year 1 – 2)

Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyderTM. 2005. Fodor et al. Bioinformatics 21(19):3733-40 (UCRL-JRNL-207079)

Statistical Analysis of the Experimental Variation in the Proteomic Characterization of Human Plasma by Two-Dimensional Difference Gel Electrophoresis. 2006. Corzett et al., J. Proteome Res. 5, 2611-2619. (UCRL-JRNL-219771)

State-based Automata Descriptions of Intracellular Protein Kinetics and Gene Regulation. 2003. J.R. Kercher. (UCRL-ID-151868)

Preliminary analysis of gene expression data from glycolysis in Yersinia pestis: Application of a prototype genetic algorithm. 2003. Kercher, J.R. et al. (UCRL-ID-152287)

Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The General Case for Multiple Groups. 2004. Kercher, J.R. et al. (UCRL-JRNL-203451)

Supplement Report on Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The Special Case of Two Groups. 2004. Kercher, J.R. et al. (UCRP-JRNL-213450)

Variable Selection in Canonical Analysis of Gene and Protein-Expression Data: The General Case for Multiple Groups. 2004. Kercher, J.R. et al. (UCRP-JRNL-205177)

Supplement to Variable Selection in Canonical Analysis of Gene- and Protein Expression Data: The General Case for Multiple Groups. Kercher, J.R. et al. (UCRP-JRNL-205176)

Aim 2. Characterize the normal range for candidate host response surrogate markers in humans and animals (year 2 – 3).

Limited Dynamic Range of Immune Response Gene Expression Observed in Healthy Blood Donors Using RT- PCR. 2006. Molecular Medicine, 12(7-8), 185 -195 (UCRL-JRNL-226594)

Aim 3. Differentiate the host response surrogate marker profile of infected sick or presyndromic animals or humans (year 2 -3).

Serum Protein Profile Alterations in Hemodialysis Patient. 2004. Nephrology 24, 268-274. (UCRL- JRNL-201081)

Early detection of infectious disease using host biochemical signatures in mice infected with cowpox virus. Langlois et al., in preparation. (UCRL-TR-226614)

Conclusions

Through this work methods have been tested that can define molecular signatures indicative of infection. This work has suggested that it is crucial that investment be made in understanding methodological variation and that effort must be invested in developing robust protocols for gel electrophoresis to be useful for biosignature discovery. However, with appropriate investment in use of well-characterized methods, it was shown that the variation in human healthy controls is limited to an approximately 3X variation around the mean suggesting that changes within one order of magnitude should be detectable. It has also been shown that in a controlled mouse study that blood-based markers can differentiate infected from healthy animals within a few days following exposure and at least a day prior to overt symptoms. Thus, it appears feasible to develop biosignatures for prodromal and potentially presyndromic infectious disease detection.

However, significant research is still needed to realize this potential. Challenges exist in discovering the proper set of individual biomakers that would comprise the signature, whether such signatures are specific for a specific infectious agent. How early these signatures can be detected and how a number of potential confounds will affect such an approach such as age, gender, previous disease history, and genetic make up.

References

Anderson, R. A. 2001. Deaths: Leading Causes for 1999, *49*

Armstrong, G. L., Conn, L. A., and Pinner, R. W. 1999. Trends in infectious disease mortality in the United States during the 20th century, Jama *281*, 61-66.

Barquet, N., and Domingo, P. 1997. Smallpox: the triumph over the most terrible of the ministers of death, Ann Intern Med *127*, 635-642.

Binder, S., and Levitt, A. M. 2002. Emerging Infectious Diseases: A Strategy for the 21st Century (Centers for Disease Control).

Cello et al., 2002. Science, 297, 1016.

De Lorenzo, R. A., and Porter, R. S. 2000. Weapons of Mass Destruction - Emergency Care (Upper Saddle River, NJ, Prentice-Hall, Inc.).

Editors. 2002. Communicable Disease Surveillance and Response (World Health Organization Regional Office for Europe).

Inglesby, T. V., Dennis, D. T., Henderson, D. A., Bartlett, J. G., Ascher, M. S., Eitzen, E., Fine, A. D., Friedlander, A. M., Hauer, J., Koerner, J. F., *et al.* 2000. Plague as a biological weapon: medical and public health management. Working Group on CivilianBiodefense, Jama *283*, 2281-2290

Kaplan, E.H., Craft, D.L., Wein, L.M., 2002. Emergency Response to a Smallpox Attack: The Case for Mass Vaccination. Proc. Natl. Acad. Sci. U.S.A., 99,10935-40.

McConnell, J. 2002. Ready for the next influenza pandemic?, Lancet *359*, 1133

Meltzer, M. I., Cox, N. J., and Fukuda, K. 1999. The economic impact of pandemic influenza in the United States: priorities for intervention, Emerg Infect Dis *5*, 659-71

Nelson, K. E., Williams, C. M., and Graham, N. M. H. 2001. Infectious Disease Epidemiology - Theory and Practice (Gaithersburg, MD, Aspen Publishers, Inc.).

Noah, D., and Fidas, G. (2000). The Global Infectious Disease Threat and Its Implications for the United States (Gordon, D. F. (National Intelligence Council)).

O'Toole, T, Maur., M., Inglesby, T.V. 2002. Shining light on "Dark Winter". Clin Infect Dis. 34, 972-83.

Snacken, R., Kendal, A. P., Haaheim, L. R., and Wood, J. M. (1999). The next influenza pandemic: lessons from Hong Kong, 1997, Emerg Infect Dis *5*, 195-203.

Appendix

Manuscripts and Data

Aim 1

Evaluate and select methods for the high throughput screening of blood for molecular signature characterization (year 1 – 2)

BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 19 2005, pages 3733–3740
doi:10.1093/bioinformatics/bti612

Gene expression

Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyder™

Imola K. Fodor[1,*], David O. Nelson[1], Michelle Alegria-Hartman[2], Kristin Robbins[2], Richard G. Langlois[2], Kenneth W. Turteltaub[2], Todd H. Corzett[2] and Sandra L. McCutchen-Maloney[2]

[1]Computation Directorate and [2]Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA

Received on June 15, 2005; revised on July 5, 2005; accepted on August 2, 2005
Advance Access publication August 9, 2005

ABSTRACT

Motivation: The DeCyder software (GE Healthcare) is the current state-of-the-art commercial product for the analysis of two-dimensional difference gel electrophoresis (2D DIGE) experiments. Analyses complementing DeCyder are suggested by incorporating recent advances from the microarray data analysis literature. A case study on the effect of smallpox vaccination is used to compare the results obtained from DeCyder with the results obtained by applying moderated *t*-tests adjusted for multiple comparisons to DeCyder output data that was additionally normalized.

Results: Application of the more stringent statistical tests applied to the normalized 2D DIGE data decreased the number of potentially differentially expressed proteins from the number obtained from DeCyder and increased the confidence in detecting differential expression in human clinical studies.

Availability: The marray and limma packages used here are available from http://www.bioconductor.org/

Contact: fodor1@llnl.gov

1 INTRODUCTION

Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) is a technology by which thousands of proteins in a biological sample are separated according to their isoelectric points and molecular weights (O'Farrell, 1975; Görg *et al*., 2000; Lilley *et al*., 2002). In theory, each protein is uniquely determined by its response along the two dimensions of separation. Differences in the proteomes of multiple samples can be studied by comparing the expression profiles of the proteins on the gels. In traditional 2D PAGE, each gel contains one sample which is compared with the samples on different gels, introducing high experimental variability.

Ünlü *et al*. (1997) proposed 2D difference gel electrophoresis (2D DIGE) as a method to overcome gel-to-gel variability inherent to 2D PAGE. More recently, 2D DIGE has been commercialized through the Ettan DIGE System of Amersham Biosciences (now a part of GE Healthcare), thanks to the development of the three size and charge-matched, spectrally resolvable CyDye fluors Cy2, Cy3 and Cy5. Gels using the DIGE method contain three samples labeled with the three distinct fluorescent dyes Cy2, Cy3 and Cy5. Typically, two dyes are used to label two different biological samples of interest. The third dye can be used to label the 'internal standard' which is a pooled mixture of all the samples used in the experiment, and is identical on all gels. The power of the internal standard is in its potential to adjust for the variability between gels and thus make the data across the experiment more comparable. The DeCyder differential analysis software is a part of the Ettan DIGE System, and is used for analyzing the data and quantifying the differential expression of the proteins (Tonge *et al*., 2001; Alban *et al*., 2003; Amersham, 2003).

Although there are fundamental differences in 2D DIGE and gene-expression microarray technologies, many of the difficulties encountered in the analysis of 2D DIGE data are similar to problems that arise in the analysis of microarray experiments: proper normalization of the data within and between the gels (arrays), multiple hypothesis testing and the quest for improved test statistics that exploit the common information across the proteins (genes) (Huber *et al*., 2002, 2003; Smyth *et al*., 2003b; Dudoit and Yang, 2003; Cui and Churchill, 2003). Since data from 2D DIGE experiments exhibit similar characteristics to microarray datasets, we adapted methods developed by researchers in the microarray field to address statistical challenges in analyzing proteomic data from 2D DIGE.

Earlier studies based on DeCyder version 4.0 proposed robust statistical methods and normalization techniques to complement the analytical tools in DeCyder (Kreil *et al*., 2004; Karp *et al*., 2004). We offer additional improvements in the assessment of differential protein expression by combining related normalization methods with novel statistical tests, based on a study with DeCyder version 5.01.

2 APPROACH

To investigate the response of the human proteome on exposure to smallpox vaccination, a proteomic study involving five human subjects, before and at five time points after vaccination, was undertaken. Based on literature indicating the advantages over other 2D gel methods (Tonge *et al*., 2001; Alban *et al*., 2003), 2D DIGE was selected as the technology platform. Blood samples were collected from five volunteers at six time points before and after vaccination, with informed consent under the Institutional Review Board approval from Lawrence Livermore National Laboratory. The samples were prepared and labeled following the manufacturer's

*To whom correspondence should be addressed.

Table 1. 2D DIGE experimental design. Each gel had three samples, two corresponding to a subject sample with time of collection indicated (labeled with Cy3 and Cy5) and a pooled standard that was common on all gels labeled with Cy2

Time	Subject S_1	S_2	S_3	S_4	S_5
T_1: 1 h prior T_2: 1 h post	Gel 1	Gel 4	Gel 7	Gel 10	Gel 13
T_3: Day 1 T_4: Day 3	Gel 2	Gel 5	Gel 8	Gel 11	Gel 14
T_5: Day 7 T_6: Day 14	Gel 3	Gel 6	Gel 9	Gel 12	Gel 15

protocol for 2D DIGE and included the removal of the six proteins with highest abundance (Chromy *et al*., 2004). Details of the sample processing are available from the authors. The resulting 30 samples were arranged on 15 gels as shown in Table 1. The 30 biological samples (five subjects, six time points) were analyzed by 2D DIGE in triplicate, resulting in 45 total gels. In two replicates, on any given gel, the sample corresponding to the earlier sampling time was labeled with Cy3, whereas the sample corresponding to the later time was labeled with Cy5. In one replicate, the dyes were swapped. All gels contained an identical third sample, the pooled standard labeled with Cy2. The scientific goal was to identify proteins that were differentially expressed in response to smallpox vaccination, as a model for smallpox. The aim of the present study was to investigate the results obtained with DeCyder and indicate possible improvements in proteomic data analysis.

DeCyder version 5.01 was used for spot detection and matching across the gels (Amersham, 2003). Both the Differential In-gel Analysis (DIA) and the Biological Variation Analysis (BVA) modules were used: the former to codetect and quantify the spots on a given gel in terms of the ratios of the Cy3 and Cy5 sample volumes to the standard Cy2 volume, and the latter to match the spots and standardize the ratios across the gels accounting for the observed differences in the Cy2 sample volumes on the gels. For each gel, the spot boundaries obtained from the Cy2 image were copied over to the images of the other two samples on the same gel. Since the internal standard was identical on all gels, the software performed the matching only on the internal standard images labeled with Cy2, without introducing sample-to-sample differences into the matching. The master gel was chosen as the gel with the most spots. The other spot maps were matched to the master image with a proprietary 'pattern recognition algorithm that matches one single spot in one gel to a single spot in another gel based on its neighboring spots' (Amersham, 2003). To increase the accuracy of the automatic gel-to-gel matching, careful manual landmarking was performed as recommended in the software documentation.

The volume of a spot for a given dye is defined as the fluorescent intensity of the corresponding dye integrated over the area of a spot. Normalized volume refers to the volume normalized across the three dyes and across the gels. One of the outputs DeCyder provides is the ratio of the normalized volumes, also called the standardized abundances,

$$\begin{cases} R_{pg} = \text{VolCy5}_{pg}/\text{VolCy2}_{pg}, \\ G_{pg} = \text{VolCy3}_{pg}/\text{VolCy2}_{pg}, \end{cases} \quad (1)$$

for each spot p and gel g in the experiment. VolCy5_{pg} represents the normalized volume of spot p on gel g in the Cy5 sample and similarly for the other two dyes.

The statistical analyses in DeCyder are based on the standardized protein log abundances, which are defined as the log10 of the standardized abundances. In theory, the standardized log abundances follow a normal distribution and are comparable across all spots and gels.

The output from DeCyder was exported and analyzed in the R computing environment (http://www.r-project.org/).

2.1 Fitting linear models to assess the differential expression of proteins

The goal of the study was to detect proteins that showed differential expression post-vaccination. Thus, all pairwise comparisons among the six time points were of interest.

DeCyder provides two choices for determining if a protein is differentially expressed between two groups: one based on the fold change and the other on the P-value from the traditional Student's t-test. Fold change is calculated as the ratio of the average standardized abundances corresponding to the two samples. If $\bar{S}_{p1}$ and $\bar{S}_{p2}$ denote the average standardized abundance of protein p in groups $i = 1$ and 2, respectively,

$$\bar{S}_{pi} = \frac{\sum_{R_{pg} \in \text{Group}_i} R_{pg} + \sum_{G_{pg} \in \text{Group}_i} G_{pg}}{|R_{pg} \in \text{Group}_i| + |G_{pg} \in \text{Group}_i|}, \quad (2)$$

then the corresponding fold change is

$$F_p = \begin{cases} +\bar{S}_{p1}/\bar{S}_{p2} & \text{for} \quad \bar{S}_{p1} > \bar{S}_{p2}, \\ -\bar{S}_{p2}/\bar{S}_{p1} & \text{for} \quad \bar{S}_{p1} < \bar{S}_{p2}. \end{cases} \quad (3)$$

A k-fold expression increase/decrease is reflected in a $+k/-k$ value of F_p; no change corresponds to $F_p = 1$.

A common way to assess the differential expression of the proteins is to combine the two measures and find the proteins that exceed a predetermined fold change with a predetermined significance.

In the microarray literature it has been shown that in order to test for the differential expression of many genes in parallel, the traditional Student's t-test can be improved upon (Cui and Churchill, 2003). One common approach is to adjust the gene-specific standard deviation estimates with adjustment factors calculated from a larger set of genes. The idea is to take advantage of the fact that the same model is fit across all genes. The detail lies in specifying how the gene-specific parameters and variances differ. Improved statistics based on empirical methods have been suggested in Baldi and Long (2001) and Efron *et al*. (2001). The moderated t-statistic introduced in Lönstedt and Speed (2002) and further explained in Smyth (2004) (http://www.bepress.com/sagmb/vol3/iss1/art3) is based on a hierarchical, hybrid classical/Bayes model and has been shown to follow a t-distribution under certain assumptions.

In addition to the traditional t-statistics, the moderated t-statistics, as implemented in Smyth *et al*. (2003a), was also used in this study in order to determine the differential expression of proteins. The problem was cast in a general linear modeling framework which facilitated testing using both methods. Consider the model

$$y_{pij} = \alpha_{pi} + \epsilon_{pij}, \quad (4)$$

where y_{pij} is the standardized log abundance of replicate j at time T_i of protein spot p, α_{pi} is the unknown expression level of protein

spot p at time T_i and ϵ_{pij} is a random error, for $p = 1, \ldots, 2384$ (number of spots), $i = 1, \ldots, 6$ (number of time points) and $j = 1, \ldots, 15$ (number of replicates at each time). To follow the analysis with DeCyder, the 3 replicates of the 5 subjects were treated as 15 replicates.

For a given spot p, let $\mathbf{y}_p$ denote the vector of the 90 observations at that spot, ordered according to time: the first 15 values are the replicates at time T_1, followed by the 15 replicates at times T_2, T_3, T_4, T_5 and T_6. Similarly, let ϵ_p denote the corresponding vector of random errors. If $\alpha_p = (\alpha_{p1}, \alpha_{p2}, \ldots, \alpha_{p6})^{\mathrm{T}}$, then the model in Equation (4) can be written in matrix terms as

$$\mathbf{y}_p = \mathbf{X}\,\alpha_p + \epsilon_p, \tag{5}$$

where the design matrix $\mathbf{X}$ has size 90×6, and its i-th column has 15 ones in its $i \times 15$th positions for $i = 1, \ldots, 6$, and is zero everywhere else.

Testing the equality of the expression levels at different times can be easily specified with appropriate contrasts, or linear combinations of the parameters. For example, testing the null hypothesis that the expression level of spot p at time T_1 is equal to the expression level at time T_2,

$$H_0 : \alpha_{p1} = \alpha_{p2}, \tag{6}$$

is equivalent to

$$H_0 : \beta_{p12} = 0, \tag{7}$$

where

$$\beta_{p12} \doteq C^{\mathrm{T}}\,\alpha_p = (-1\ 1\ 0\ 0\ 0\ 0)\,\alpha_p. \tag{8}$$

For each spot in the experiment, the 15 pairwise comparisons among the six time groups were performed, using both the traditional (corresponding to the results from DeCyder) and the moderated t-statistics.

2.2 Normalizing the standardized log abundances

The distribution of the standardized log abundances showed systematic biases within the gels and had different ranges across the gels. Since both of these problems have been encountered by the microarray analysis community, methods developed to address these issues in microarrays were investigated. Specifically, the limma Norm package from the Bioconductor project (Smyth *et al*., 2003a) was used.

To perform the additional normalizations, the standardized abundances in Equation (1) were first transformed into the $M - A$ space, where

$$\begin{cases} M_{pg} = \log_2(R_{pg}/G_{pg}), \\ A_{pg} = 1/2 \log_2(R_{pg} \times G_{pg}). \end{cases} \tag{9}$$

A_{pg} measures the Average, and M_{pg} (Minus) the difference between the intensities of the two samples (samples labeled with Cy3 and Cy5, respectively) on a log scale at spot p on gel g. Assuming that the majority of the proteins were not differentially expressed between the two conditions, the plot of M_{pg} versus A_{pg} (MvA) for a given gel should result in a random scatter around the zero-line with no systematic trends. Observed systematic variations may be the result of different labeling efficiencies for the Cy3 and Cy5 dyes, as well as different scanning settings and gel effects. In microarrays, dye imbalances often vary according to the average spot intensity A (Smyth *et al*., 2003b). The MvA plots for the 45 gels exhibited systematic trends which depended on the value of A (Fig. 4a and 4b); therefore, local intensity-dependent regression lines through the data were fitted using the loessFit function in R. Next, the M-values were replaced by the residuals from the fit which resulted in pattern-free MvA plots (Fig. 4c and 4d). The second normalization step used boxplots for between-gel normalization (Fig. 5). It involved comparing the ranges of the regression-corrected M-values across the 45 gels, and scaling them so that the middle 50% of the data on each gel spanned the same range.

Let $\tilde{M}_{pg}$ and $\tilde{A}_{pg}$ denote the corrected values after the MvA normalization within gels and boxplot normalization between gels. Next, the inverse transformation of Equation (9) was used to transform $\tilde{M}_{pg}$ and $\tilde{A}_{pg}$ back to the original RG scale, and obtain the normalized standardized abundances $\tilde{R}_{pg}$ and $\tilde{G}_{pg}$ corresponding to Equation (1). The standardized abundances from DeCyder were thus further normalized.

The linear model fitting described in Section 2.1 was repeated at each of the spots, using the log10 of $\tilde{R}_{pg}$ and $\tilde{G}_{pg}$ as the response variable in Equation (4). The model was identical to Equation (5), except that the data at each spot consisted of the 90 normalized standardized log abundances instead of the 90 standardized log abundances.

2.3 Adjusting the P-values

Another challenge in the analysis of 2D DIGE data that is shared with the microarray data analysis community is the massive multiple hypothesis problem (Shaffer, 1995). Regardless of the data used and the testing procedure employed, the resulting P-values need to be adjusted because numerous tests are performed simultaneously. The unadjusted P-values that result from the individual t-tests applied separately at each time point pair and at each spot are too optimistic. At the $\alpha = 0.05$ significance level, 1 every 20 tests is expected to result in a P-value less than α just by chance. As the number of tests increases, so does the number of false positives. Several adjustment methods have been proposed. The simplest one is the Bonferroni correction, which multiplies the unadjusted P-values by the total number of tests performed. A less stringent, but more practical approach for the present case is the false discovery rate method of Benjamini and Hochberg (1995). Let R denote the total number of rejected hypotheses, and V the number of falsely rejected hypotheses, out from the total number of simultaneous tests. Then, the realized False Discovery Rate (FDR) is defined as V/R, for $R > 0$, and 0 otherwise. Since V is unobserved, Benjamini and Hochberg (1995) developed a sequential P-value procedure that controls the *expected* value of the FDR, E(FDR), under the assumption that the test statistics are independent. The resulting process controls E(FDR) at the fixed level α for any joint distribution of the P-values. Although the independence assumption is not always satisfied, the FDR method is often used because of its simplicity. Since its results are preferable over the unadjusted P-values, here the FDR procedure in R was used.

3 RESULTS

Figure 1 displays the standardized log abundance data for one protein spot. Assuming that a protein was present in all the samples and that its corresponding spot was found and matched across all 45 gels, there should be 15 values at each time point: three replicates for each of the five subjects. For spot 1186, the third replicate of gel 8

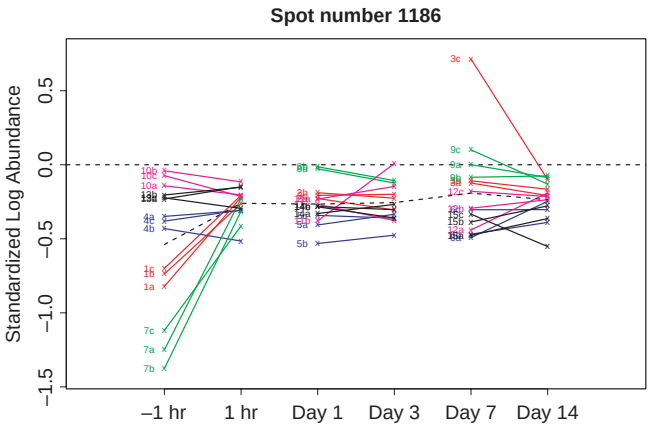


Fig. 1. The standardized log abundance for one spot. Numbers indicate gels, letters stand for replicates, and colors represent subjects. The dotted line connects the averages at the six time points.

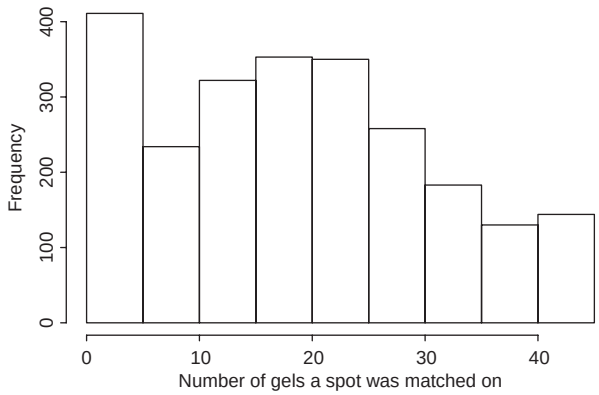


Fig. 2. Histogram of the number of gels a spot was matched on: 2384 spots and 45 gels.

is missing, evidenced by the two green lines connecting Day 1 and Day 3 in Figure 1.

A total of 2384 spots were identified on the master gel, defined to be the gel containing the most spots. Figure 2 presents the histogram of the number of gels a spot was matched on. Fewer than 150 spots were matched on at least 40 of the 45 gels. The less stringent criterion requiring at least five observations at each time point resulted in 1026 spots.

3.1 Results with the Student's *t*-statistic using the standardized log abundances

Table 2 presents the number of spots with >1.5-fold change, and with P-value <0.05, for each of the 15 pairwise comparisons involving the data at two time points. The response was the standardized log abundance and the test was based on the traditional t-statistics. The values in the unadjusted columns used the unadjusted P-values that resulted from performing the traditional t-tests independently at each of the spots and time pairs. The fold changes and the P-values corresponding to the individual spots under the unadjusted heading match the results given by DeCyder. The FDR-adjusted columns refer to P-values that were adjusted for the multiple comparisons. Comparing

Table 2. The number of spots with >1.5-fold change and P-value ≤0.05. Pairwise tests using the standardized log abundances and Student's t-test

	Unadjusted					FDR-adjusted				
	T_2	T_3	T_4	T_5	T_6	T_2	T_3	T_4	T_5	T_6
T_1	7	47	62	53	54	0	8	11	8	11
T_2		47	53	71	59		11	15	8	11
T_3			3	32	49			1	5	13
T_4				55	58				9	19
T_5					8					1

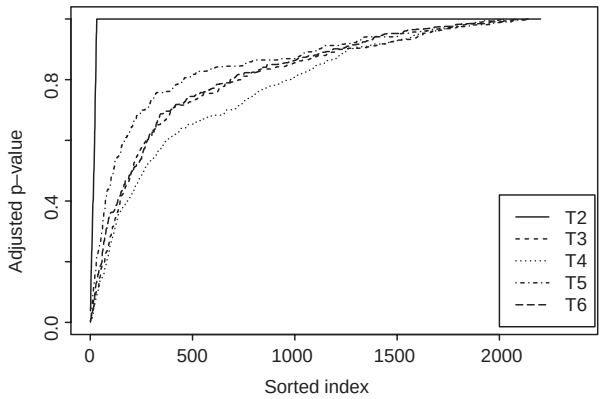


Fig. 3. Sorted FDR-adjusted P-values for the pairwise t-tests that compare the average standardized log abundances at time T_1 to the subsequent time points.

the corresponding numbers under the unadjusted and FDR-adjusted cells in Table 2 illustrates the effect of adjusting for multiple comparisons. The number of 'interesting' spots decreases dramatically after the multiple statistical hypothesis testing problem is addressed.

When aggregating the possibly overlapping results of the 15 pairwise comparisons, a total of 310 unique spots had >1.5-fold change and unadjusted P-value <0.05 in at least one pairwise test. The corresponding number based on the FDR-adjusted P-values was 83.

Figure 3 displays the sorted adjusted P-values from the pairwise t-tests calculated at each spot comparing the five subsequent times to T_1. A possible explanation for the unique shape of the T2 versus T1 curve (solid) compared with the other curves in Figure 3 is the fact that the T2 versus T1 comparisons involved spots from the same gels, whereas the others compared spots from different gels. Example statistics for the number of spots included in the intragel versus intergel comparisons for Subject 1 were: T2 versus T1: 1133 spots (equal to the number of spots on gel 1a that were matched with the spots on the master gel), T3 versus T1: 714 spots (the number of spots on gel 1a that were matched with the spots on both gel 2a and the master gel), T4 versus T1: 714 (same as for T3 versus T1), T5 versus T1: 780 spots (the number of spots on gel 1a that were matched with the spots on both gel 3a and the master gel), T6 versus T1: 780 (same as for T5 versus T1). Similar trends existed for the other subjects as well: more (and better matched spots) for intragel comparisons, fewer (and less well matched) spots for intergel comparisons.

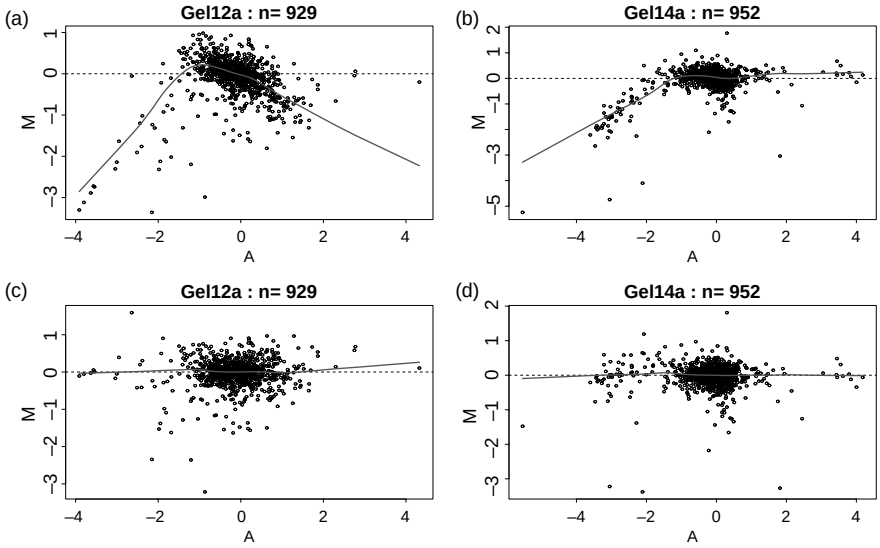


Fig. 4. The MvA plots for gels 12a and 14a: (**a**) and (**b**) based on the standardized log abundances from DeCyder, (**c**) and (**d**) the corresponding results after the loess normalization. The titles reflect the number of spots from the given gel matched to spots on the master gel.

Table 3. The number of spots with >1.5-fold change and FDR-adjusted P-val $\leq$0.05. Pairwise tests using the moderated t-statistics and (a) the standardized log abundances and (b) the normalized standardized log abundances.

	(a)					(b)				
	T_2	T_3	T_4	T_5	T_6	T_2	T_3	T_4	T_5	T_6
T_1	1	3	4	3	0	1	4	4	0	0
T_2		4	9	5	2		7	7	5	6
T_3			1	2	2			0	4	2
T_4				3	2				4	4
T_5					1					0

3.2 Results with the moderated t-statistic using the standardized log abundances

Panel (a) of Table 3 is similar to the FDR-adjusted panel of Table 2, and presents the corresponding results obtained using the moderated t-statistic along with the standardized log abundances. Results with the unadjusted P-values were generally higher, but overall comparable to the unadjusted results in Table 2. Aggregating the results of the FDR-adjusted P-values from panel (a) of Table 3 from all 15 pairwise tests resulted in 13 unique spots.

3.3 Results with the moderated t-statistic using the normalized standardized log abundances

Figure 4 displays MvA plots for two gels, before (a, b) and after (c, d) the normalizations within the gels. The data for most of the other gels showed similar characteristics. Figure 5 shows the effect of the additional between-gel normalization step. Figure 5a displays the boxplots of the M values based on the output from DeCyder. Differences among the gels are clearly visible, especially for gel 3c which had a higher interquartile range (the middle 50% of the data values within the boxes of the boxplots) than any of the other gels. The unusual distribution for gel 3c was probably caused by problems specific to either that gel or the processing of that gel, as the corresponding distributions for replicates 3a and 3b did not exhibit such anomalies. Figure 5b presents the corresponding results after within-gel normalization. Consequent to the local regression fit, the boxplots in Figure 5b are all centered around zero. However, the interquartile ranges show differences across the gels. The between-gel normalization step brings the interquartile ranges of the gels onto the same scale, as shown in Figure 5c. After the MvA normalization within arrays and boxplot normalization between arrays, the normalized standardized log abundances corresponding to the six time points in the experiment were obtained as described in Section 2.2. Figure 6 displays the result for spot 1186 whose standardized log abundance data were shown in Figure 1.

Panel (b) of Table 3 presents the number of spots with a >1.5-fold change and FDR-adjusted P-value $\leq$0.05, using the normalized standardized log abundances as the response variable and testing with the moderated t-statistics. Combining the results of the 15 pairwise tests resulted in 13 unique spots. Results with the unadjusted P-values were generally higher, but overall comparable to the unadjusted results in Table 2.

4 DISCUSSION

Figure 7 aggregates the results of the three FDR-adjusted methods in Section 3 in a Venn diagram. The numbers in the circles represent unique spots. Of the eight spots commonly identified by all three adjusted methods, only one spot (2196) had enough observations to be of practical interest from a statistical perspective, loosely defined here as having at least five observations at each time, irrespective of which subject the available replicates belonged to and keeping in mind that subject variability and host response could result in differential expression. Of the three spots commonly identified by TAdjs and NormModTAdj, two (1506 and 1596) contained the required

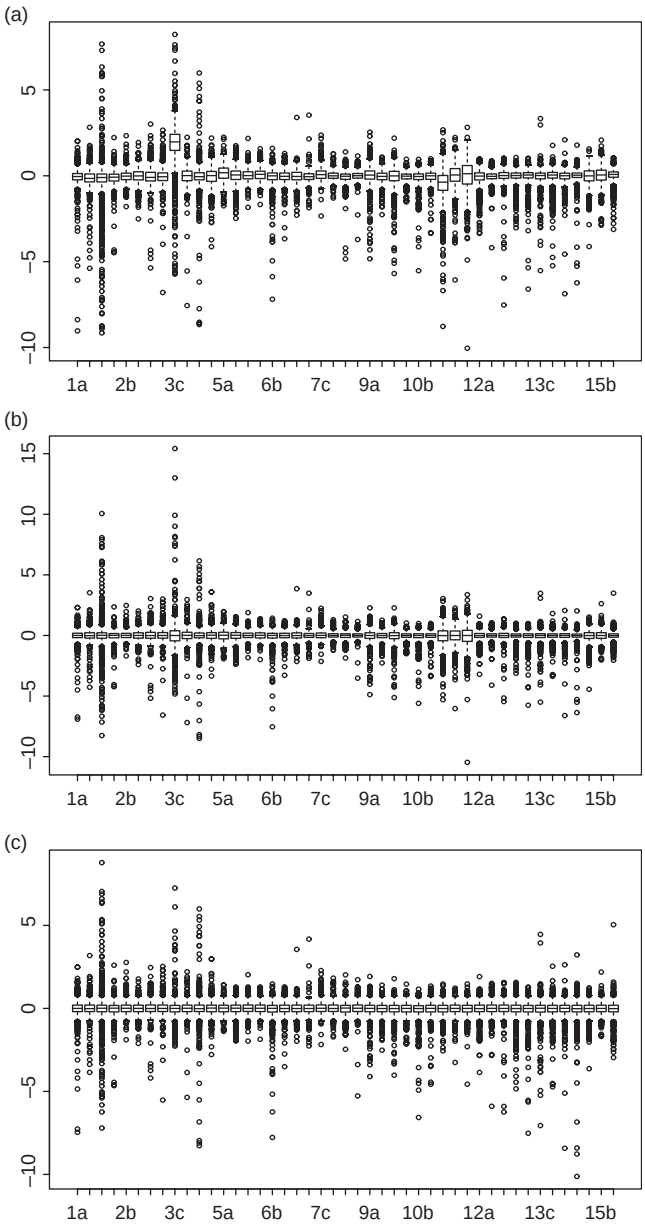


Fig. 5. The boxplots of the M-values for the 45 gels (**a**) before, (**b**) after within-gels and (**c**) after within- and between-gel normalization. The gels are ordered sequentially according to the experimental design in Table 1: the three replicates of gel 1 (1a, 1b, 1c) followed by the three replicates of gels 2 through 15 (15a, 15b, 15c).

number of data points. Being identified by more than one adjusted method suggests a higher confidence that these spots represent proteins that are indeed differentially expressed. Confirmation requires protein identification by mass spectrometry followed by further validation experiments. The three spots identified only by the ModTAdj method, the two spots identified only by the NormModTAdj method and the two spots commonly identified by TAdjs and ModTAdj, each had less than five values per time point, so in this case were not considered although important information may still be found from these patterns.

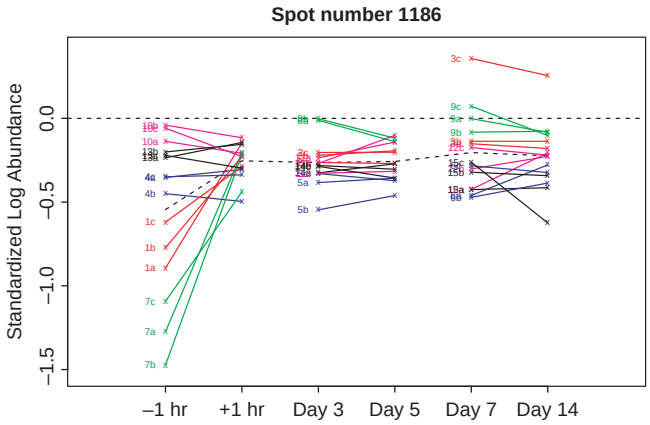


Fig. 6. The normalized standardized log abundance data corresponding to Figure 1.

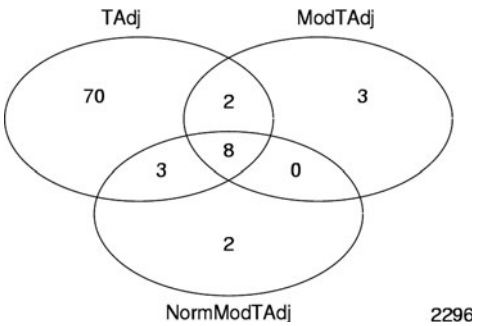


Fig. 7. Venn diagram comparing the results based on the three FDR-adjusted methods in Table 2 (TAdj), Panels (a) (ModTAdj) and (b) (NormModTAdj) of Table 3.

Several factors contributed to the higher complexity of this clinical study, as compared with other published 2D DIGE experiments: (1) the choice of using human blood, one of the most complex proteomes with estimates of 100 000 circulating proteins with a wide dynamic range in concentrations; (2) subject-to-subject variability within the five vaccinees; (3) challenges of variable host immune response; (4) the large number of gels involved. In addition, the gels were prepared in-house. Although the extent of the following challenges is expected to be less severe in simpler experiments, the qualitative conclusions drawn here remain valid for other 2D DIGE studies as well. Our preliminary findings with precast gels (whose reproducibility has been improving in recent years) suggest significant improvements in the quality of the data.

4.1 Normalization

We found evidence for inadequate normalization of the data within and between the gels. Our results agree with other recent findings (Kreil *et al.*, 2004; Karp *et al.*, 2004), and indicate the need to develop better techniques. Since the global characteristics of the data resembled data from microarray experiments, we suggested methods developed in that community as possible ways to improve the normalization of proteomic data from 2D DIGE.

4.2 Accounting for multiple comparisons

Whenever there are multiple hypothesis tests, the observed significance levels have to be adjusted. Here the FDR method was used.

4.3 Matching spots across gels

Although the spot matching rates observed in this study may seem low, there are no reports upon which to compare our results for a human plasma clinical study. Published studies citing 52% (Alban *et al*., 2003) and 67% (Yan *et al*., 2002) of spots matched on gels relied on far fewer gels (12 and 8, respectively) and the use of simpler biological samples (*Escherichia coli*) which would not be affected by genetic variability characteristic to human subjects. In addition, differences in spot matching can be attributed to the wide isoelectric point (pI) and molecular weight (mw) region used in our study: non-linear pI range 3–10, mw range 200–20 kDa. By targeting a narrower pI or mw region, protein spots would be better resolved with improved subsequent matching results. The number of spots specified as an input to the DeCyder algorithm also affects the results. The strategy in this study was to start with a large initial spot number (2500) in order to maximize detection of small-abundance proteins. The large number of spots specified, however, could lead to the inclusion of dust particles or other artifacts. Thus, the current state of the technology is not fully automated, and all potentially interesting spots should be manually verified.

4.4 Spot migration

Microarrays consist of a fixed grid of spots, where each spot contains a unique DNA sequence from a known gene. In contrast, proteins migrate through the gels according to their pI and mw. Genetic differences between subjects and post-translational modifications may result in certain protein spots missing from certain gels, or the 'same' protein migrating slightly differently on the gels. The challenge is to untangle the biological differences in protein expression from differences owing to experimental variation. Spot migration is thus one fundamental difference between microarrays and gels that needs to be addressed, in particular as it relates to spot matching and model development. The mechanistic approach of this paper to ignore spots with poor matching was only a first attempt to understand the data. More sophisticated methods that take into account the underlying biology should be developed, as unmatched spots between subjects may hold information of biological interest.

4.5 Intragel versus intergel comparisons

Although the internal standard is used in 2D DIGE to guarantee that all spots are comparable across all gels, we found evidence to the contrary. The distinct shape of the T2 versus T1 curve, compared with all other time points in Figure 3, points to the different nature of comparing samples from the same gel and comparing samples from different gels. Such differences are most likely because of the imperfect intergel matching. The distinct pattern of the T2 versus T1 curve persisted over the T4 versus T3 and the T6 versus T5 comparisons, but not over the other pairwise comparisons. To minimize the effects of matching, samples of most interest in comparing should be placed on the same gel. Improvements in spot detection and matching should mitigate the differential effects observed in the intergel comparisons. Performing the spot detection separately on each gel image (instead of only on the Cy2 images) may increase the accuracy. The high complexity of the internal standard may have contributed to the poor matching. Perhaps a simpler internal standard consisting of all the T1 samples, or including on all gels an identical T1 reference sample labeled with either Cy3 or Cy5, would have led to superior results. These and other alternatives should be explored, balancing the cost of running the experiment with the quality of the results.

4.6 Statistical modeling

Proper experimental design should be an integral part of any experiment. The design in Table 1 was chosen following recommendations in Amersham (2003). To formulate the optimal design for a given experiment, we advocate interaction with statisticians on the allocation of the samples to the gels, and on proper randomization. Results for microarrays (Kerr and Churchill, 2001) could be extended.

The linear modeling framework of Smyth *et al*. (2003a) used here provides a flexible extension to the simple tests provided in DeCyder. Testing additional hypotheses involving different subsets of the subjects and the time points amounts to specifying different design matrices and contrasts, then proceeding with the estimation as described within. Functionality in R allows one to fit the linear models using robust techniques that minimize the effects of outliers. Accounting for the different number of data points at the different spots is automatically included in the models.

Although the moderated t-test provides an alternative to the Student's t-test for pairwise comparisons, other methods are also possible. From a statistical perspective, a more appropriate way to analyze the data is to fit a mixed effect model at each spot, treating the subjects as five blocks and the gels as two blocks within the subjects (Pinheiro and Bates, 2000). Then, one test at each spot is used to determine if there are any differences among the six time points. Including the block effects improves the estimation of the time effects of interest, and separates the biological replicates from the technical replicates. The two-factor Analysis of Variance (ANOVA) model in DeCyder only supports fixed effects, and is unable to model the random subject and gel effects. Since both the subjects and the gels are samples from larger populations, random effects are appropriate for them. We performed the described mixed-effect modeling at each spot, and found four spots with FDR-adjusted P-value for a time effect <0.05 and at least a 1.5-fold change between any two time points. Of the four spots, one spot (2196) was previously selected by all three adjusted methods. Since spot 2196 was identified by a number of different methods, it has the highest confidence that it is indeed an example of a differentially expressed protein following smallpox vaccination.

The statistical models used here have certain assumptions, such as normality of the errors and independence of the observations. However, these models can be used in an exploratory fashion even if the data exhibit departures from the assumptions (Smyth, 2004). Further model developments should incorporate more realistic assumptions about the data. In addition, they should also take into account the state of the proteins, which will require close collaboration between the proteomics and statistics communities.

5 CONCLUSION

The 2D DIGE technology plays an important role in proteomics, and rigorous data analysis techniques are essential in quantifying the differential expression of proteins between biological samples. Here, we presented readily available statistical methods to improve the analysis of 2D DIGE experiments. Our goal was to offer analytical improvements with small investment to the user. We achieved this

goal by borrowing methods from the microarray literature, and showing their feasibility and suitability to the analysis of 2D gels. To objectively quantify the effects of the proposed techniques, we are currently undertaking a technical variability study using human blood samples.

In addition to the problems shared with microarrays, 2D DIGE presents additional difficulties in spot detection and matching, especially when used in complex studies involving clinical plasma samples. Future advances in image processing and in statistical modeling specific to proteomics will further enhance the quality of 2D DIGE results. Version 6.0 of DeCyder, released after the completion of this study, offers improvements over the version used here in areas such as normalization and adjusting the significance levels in multiple comparisons. We will take full advantage of the latest software in the future.

ACKNOWLEDGEMENTS

We wish to acknowledge our clinical collaborators Harry Lampiris and Lynn Pulliam from the San Francisco Veterans Affairs Medical Center for their assistance with this study. This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, with support from the Department of Homeland Security (Biological Countermeasures Program). This work was supported by Laboratory Directed Research and Development funding. UCRL-JRNL-207079.

Conflict of Interest: none declared.

REFERENCES

Alban,A. *et al*. (2003) A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics*, **3**, 36–44.

Amersham (2003) *DeCyder Differential Analysis Software User Manual, Version 5.0*. Amersham Biosciences.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences in gene changes. *Bioinformatics*, **17**, 509–519.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statist. Soc. B*, **57**, 289–300.

Chromy,B.A. *et al*. (2004) Proteomic analysis of human serum by two-dimensional differential gel electrophoresis after depletion of high-abundant proteins. *J. Proteome Res.*, **3**, 1120–1127.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA experiments. *Genome Biol.*, **4**, 210.

Dudoit,S. and Yang,Y.H. (2003) Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. *The Analysis of Gene Expression Data: Methods and Software*. Springer, NY, pp. 73–101.

Efron,B. *et al*. (2001) Emprirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Görg,A. *et al*. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **21**, 1037–1053.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.

Huber,W., von Heydebreck,A. and Vingron,M. (2003) Analysis of Microarray Gene Expression Data. *Handbook of Statistical Genetics*, 2nd edn. Wiley, Vol 1, 162–187.

Karp,N.A. *et al*. (2004) Determining a significant change in protein expression with DeCyder during a pair-wise comparison using two-dimensional difference gel electrophoresis. *Proteomics*, **4**, 1421–1432.

Kerr,K. and Churchill,G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Kreil,D.P. *et al*. (2004) DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. *Bioinformatics*, **20**, 2026–2034.

Lilley,K.S. *et al*. (2002) Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Curr. Opin. Chem. Biol.*, **6**, 46–50.

Lönstedt,I. and Speed,T.P. (2002) Replicated microarray data. *Stat. Sin.*, **12**, 31–46.

O'Farrell,P.H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.*, **250**, 4007–4021.

Pinheiro,J.C. and Bates,D.M. (2000) Statistics and Computing. *Mixed-Effects Models in S and S-PLUS*. Springer, NY, pp. 8–11.

Shaffer,J.P. (1995) Multiple hypothesis testing. *Ann. Rev. Psych.*, **46**, 561–576.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**.

Smyth,G.K., Thorne,N. and Wettenhall,J. (2003a) *LIMMA: Linear Models for Microarray Data User's Guide*. The Walter and Eliza Hall Institute of Medical Reseach.

Smyth,G.K., Yang,Y.H. and Speed,T. (2003b) Statistical Issues in cDNA Microarray Data Analysis. *Methods in Molecular Biology*, Humana Press, Totowa, NJ, Vol. 224, pp. 111–136.

Tonge,R. *et al*. (2001) Validation and development of fluorescence two-dimensional gel electrophoresis proteomics technology. *Proteomics*, **1**, 377–396.

Ünlü,M. *et al*. (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, **18**, 2071–2077.

Yan,J.X. *et al*. (2002) Flourescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*. *Proteomics*, **2**, 1682–1698.

research articles Journal of proteome research

Statistical Analysis of the Experimental Variation in the Proteomic Characterization of Human Plasma by Two-Dimensional Difference Gel Electrophoresis

Todd H. Corzett,§ Imola K. Fodor,§ Megan W. Choi, Vicki L. Walsworth, Brett A. Chromy, Kenneth W. Turteltaub, and Sandra L. McCutchen-Maloney*

Biosciences Directorate, Lawrence Livermore National Laboratory, 7000 East Avenue, L-452, Livermore, California 94550

Received March 16, 2006

The complexity of human plasma presents a number of challenges to the efficient and reproducible proteomic analysis of differential expression in response to disease. Before individual variation and disease-specific protein biomarkers can be identified from human plasma, the experimental variability inherent in the protein separation and detection techniques must be quantified. We report on the variation found in two-dimensional difference gel electrophoresis (2-D DIGE) analysis of human plasma. Eight aliquots of a human plasma sample were subjected to top-6 highest abundant protein depletion and were subsequently analyzed in triplicate for a total of 24 DIGE samples on 12 gels. Spot-wise standard deviation estimates indicated that fold changes greater than 2 can be detected with a manageable number of replicates in simple ANOVA experiments with human plasma. Mixed-effects statistical modeling quantified the effect of the dyes, and segregated the spot-wise variance into components of sample preparation, gel-to-gel differences, and random error. The gel-to-gel component was found to be the largest source of variation, followed by the sample preparation step. An improved protocol for the depletion of the top-6 high-abundance proteins is suggested, which, along with the use of statistical modeling and future improvements in gel quality and image processing, can further reduce the variation and increase the efficiency of 2-D DIGE proteomic analysis of human plasma.

Keywords: 2-D DIGE • human plasma • proteomics • statistical analysis • technical variation • variance decomposition

Introduction

While plasma is a valuable specimen for biomarker discovery, it is one of the most complex proteomes known.[1] The large number of proteins with concentration ranges differing by more than 10 orders of magnitude, the variation within and between individuals, and the lack of universally adopted sample processing methods render biomarker discovery from human plasma extremely challenging. A number of research groups are currently working on addressing some of these obstacles, including the Human Plasma Proteome (HPP) Project of the Human Proteome Organization (HUPO).[2,3] Although a number of advances have been made by the HPP Project,[4] poor reproducibility has made it difficult to identify proteins of potential use as disease-specific biomarkers. For example, agreement in protein identification on repeat analysis of the same specimen in the same lab has been reported[5] to be less than 50%. Clearly, the low reproducibility of the results presents a significant roadblock to the practical implementation of biomarker discovery. Before individual variation can be understood and disease-specific markers can be identified from human plasma, it is necessary to quantify the sources of variation inherent in the proteomic methods and to improve them to the extent possible.

Two-dimensional difference gel electrophoresis (2-D DIGE) is a type of polyacrylamide gel electrophoresis (PAGE), which separates proteins in a sample according to their isoelectric points (p*I*) and molecular weights (MW).[6,7] In traditional PAGE, gels contain only one sample, which requires comparison across gels to discern differences in protein mobility and quantity characteristic to proteomes from multiple samples. Inhomogeneities in the gels and inconsistent staining have been reported to produce high experimental variation. 2-D DIGE was proposed as a method to overcome gel-to-gel variation inherent in PAGE,[8] and has been commercialized through the Ettan DIGE System of Amersham Biosciences (now part of GE Healthcare) with implementation of three size and charge-matched, spectrally resolvable CyDye fluors Cy2, Cy3, and Cy5.[9−12] Gels using the DIGE method contain three samples labeled with the three distinct fluorescent dyes Cy2, Cy3, and Cy5. Typically, two dyes are used to label two different biological samples of interest. The third dye is used to label an "internal standard", which is a pooled mixture of all the samples within an experiment, and is identical on all gels. The role of the internal standard is to adjust for the variability

* To whom correspondence should be addressed. Sandra L. McCutchen-Maloney. Tel: 925-423-5065. Fax: 925-422-2282. E-mail: smaloney@llnl.gov.
§ Contributed equally.

10.1021/pr060100p CCC: $33.50

Published on Web 09/20/2006

between gels and thus make the data across the gels more comparable. The DIGE system was demonstrated and validated with mouse liver homogenates[9] and *Escherichia coli*.[10] More recent results investigated the technical variation of DIGE using mouse brain, heart, and liver tissues, as well as *Erwinia caratova* bacterial samples.[13] However, to date, no published results exist on the variability of 2-D DIGE analysis of human plasma. This study reports on the experimental variation in the 2-D DIGE procedure with a human plasma sample independently prepared eight times and analyzed in triplicate. The study followed established protocols for 2-D DIGE, including the depletion of the top-6 high-abundance proteins.[14] The goal was to quantify the components of variation and to establish appropriate baseline variation estimates, which can guide the experimental design for 2-D DIGE proteomic studies with human plasma.

Materials and Methods

Sample Collection. A blood sample was taken from a healthy volunteer with informed consent under Institutional Review Board approval from Lawrence Livermore National Laboratory. The sample was collected in a 5 mL BD Vacutainer Plasma Preparation Tube (BD Biosciences, Franklin Lakes, NJ), gently inverted 10 times, and stored upright at 4 °C. Plasma was isolated from the whole blood by centrifugation at 1100 RCF at room temperature for 10 min. The separated plasma was divided into eight aliquots labeled A−H, and stored at −80 °C until further analysis.

Top-6 High-Abundance Protein Depletion. The eight aliquots of plasma were processed to deplete the top-6 high-abundance proteins[15] using the Agilent Multiple Affinity Removal System (Agilent Technologies, Palo Alto, CA) with previously published protocols.[14] Briefly, 40 μL of each sample was combined with 160 μL of Agilent buffer A in a 0.22 μm spin tube, and filtered by centrifugation for 1 min at 12 000 rpm at room temperature. Samples were loaded into a Shimadzu injector tube, and 150 μL was autoinjected into the 4.6 × 100 mm column at room temperature of a Shimadzu VP HPLC system. The plasma fraction with the depleted top-6 high-abundance proteins was collected between 2 and 4 min using a flow rate of 0.5 mL/min with Agilent buffer A. After 10 min, the top-6 fraction was eluted with Agilent buffer B using a 1 mL/min flow rate. Top-6 fractions were collected between 13 and 14.5 min, and the column was regenerated with Agilent buffer A before the injection of the next sample. Before the first plasma sample was injected an equilibration was performed with a blank injection of 160 μL of Agilent buffer A, and the spectra (280 and 590 nm) were checked for a proper baseline. Multiple runs of each sample were required to obtain adequate amounts of protein.

Protein Sample Cleanup and Protein Assay. One milliliter of the top-6-depleted protein fractions was cleaned using the PhaseOne 2-D Clean-Up kit (GE Healthcare) following the manufacturer's recommendations. The protein pellets were resuspended in 75 μL of labeling buffer containing 7 M urea, 2 M thiourea, 20 mM Tris, and 4% CHAPS, pH 8.5. Multiple preparations from the top-6 depletion for each sample were pooled after cleanup, and the protein concentration of each sample was determined using the ADV01 advanced protein assay (Cytoskeleton, Denver, CO).

2-D DIGE and Gel Imaging. To assess the variability of the 2-D DIGE system, the eight top-6-depleted plasma samples labeled A−H were randomized in triplicate in a 12-gel experiment (Table 1). Experimental replicates were included, similar to the biological replicates recommended by GE Healthcare, to maximize the likelihood of detecting any sample-to-sample variation. An internal pooled standard consisting of an equal amount of each of the eight samples was labeled with the Cy2 dye (GE Healthcare) and run on each gel. Each sample was dye-swapped and labeled with both the Cy3 and the Cy5 dyes (GE Healthcare) to investigate dye-to-dye variations.

Table 1. 2-D DIGE Experimental Design

gel number	Cy3	Cy5	Cy2
1	A	F	standard
2	G	B	standard
3	E	H	standard
4	D	A	standard
5	F	D	standard
6	C	G	standard
7	E	B	standard
8	A	C	standard
9	H	D	standard
10	C	F	standard
11	B	H	standard
12	G	E	standard

Following the manufacturer's recommended protocol, 50 μg of each sample was minimally labeled[12] with 400 pmol of the appropriate Cyanine dye and pooled for each analytical gel. One preparative pick gel (to be used for isolation ofthe proteins of interest for identification) was loaded with an additional 200 μg of unlabeled protein consisting of an equal amount of each sample, to ensure an adequate amount of protein for identification by mass spectrometry (MS). Protein samples were adjusted to a total volume of 450 μL with rehydration buffer and loaded onto 24 cm, pH 3−10, nonlinear Immobiline DryStrips and IPG buffer (GE Healthcare) for first dimension separation. Isoelectric focusing was carried out using the Ettan IPGphor II (GE Healthcare) as follows: 30 V rehydration for 12 h, 500 V for 1 h, 1000 V for 1 h, and 8000 V for 62 500 Vh. The IPG strips were then conditioned for 15 min in equilibration buffer containing 2% SDS, 50 mM Tris-HCl, pH 8.8, 6 M urea, 30% glycerol, 0.002% bromophenol blue, and 10 mg/mL DTT. After conditioning, the strips were alkylated for 15 min with equilibration buffer, but with 25 mg/mL iodoacetamide replacing the DTT. The strips were then loaded onto 26 cm × 20 cm precast 12.5% Tris-glycine polyacrylamide gels (Jule Inc., Milford, CT) and run at 2 W/gel constant power at 22 °C, using an Ettan DALT 12 (GE Healthcare), until the bromophenol blue dye-front reached the end of the gels.

Gels were scanned using a Typhoon 9410 imager (GE Healthcare) with a 100 μm resolution and adjusted PMT values to optimize sensitivity, yet prevent oversaturation. The Cy2 dye was excited at 488 nm, and emission spectra were obtained at 510 nm; the Cy3 dye was excited at 550 nm, and emission spectra were obtained at 570 nm; and the Cy5 dye was excited at 650 nm, and emission spectra were obtained at 670 nm. All gel images were cropped to the same size using ImageQuant v5.2 (GE Healthcare) to remove the edges of the gels.

Data Analysis. The DeCyder Differential Analysis Software v5.01 (GE Healthcare) was used for quantifying the differential expression of the proteins. The Differential In-gel Analysis (DIA) module was used to determine the optimal spot detection settings. Images were loaded into the Batch Processor module with the estimated number of spots set to 2500. The estimated number of spots was selected to maximize proteomic charac-

terization of human plasma based on previous experiments. Choosing a greater estimated number of spots tends to split large protein spots and increase detection of artifacts, while fewer estimated number of spots excludes less abundant proteins. The master gel was assigned automatically to the gel with the most spots detected. Each sample was grouped for analysis in the Biological Variation Analysis (BVA) module. During batch processing, the Cy2 channel from each gel was used for normalization of the spot intensities and for automated matching between gels. For each spot on each gel, the software reported the standardized abundance (SA) as the ratio of the volume in the Cy3 (or Cy5) sample to the volume of the pooled standard sample labeled with Cy2, where the volumes have been normalized across the gels. Standardized log abundance (SLA), defined as log(SA), was used in quantifying differential expression. All possible pairwise comparisons were made to detect sample-to-sample (samples A−H), dye-to-dye (Cy3−Cy5), and gel-to-gel (gels 1−12) variation. Within the BVA module, each comparison was filtered to find the spots (a) having a p-value ≤ 0.05 for the paired T-test testing the equality of the average SLA in the two groups under consideration, (b) having a greater than 1.5-fold change in expression between the groups, and (c) being correctly matched in at least two-thirds of the gels. Fold change was calculated as the ratio of the average SA in the two groups. If R denotes that ratio, the fold change F was defined as $F = R$ if $R \geq 1$, and $F = -1/R$ otherwise. A k-fold expression increase/decrease corresponded to a $+k/-k$ value of F. The analysis was converted into DeCyder 2-D (v6.5), and the Extended Data Analysis (EDA) module (GE Healthcare) was used to perform expression pattern clustering.

Spot characteristics calculated by DeCyder were exported for further statistical processing in the R statistical computing environment[16] (http://www.r-project.org/). Summary statistics were calculated for the distributions of the SLA of all spots within a gel, separately for the 12 gels. On the basis of these distributions, 95% prediction intervals for the expression ratios in future experiments were constructed.[9]

Spot-wise summary statistics, such as the coefficient of variation (CV) for the SA and the standard deviation (SD) for the SLA, were also calculated. The CV values indicated the consistency of the method, while the SD values permitted sample size and power calculations for future proteomic experiments involving human plasma. In a 1-factor ANOVA experiment, for example, the number of replicates required to detect a predetermined difference between the treatment means with a given significance level (α = Type I error in statistical hypothesis testing, i.e. the probability of a false positive, or erroneously detecting differential expression when there is none) and power (1-Type II error $= 1 - \beta$, where β is the probability of a false negative, or the probability of not detecting differential expression when in fact there is) can be determined using the noncentral F distribution, provided that an estimate of the error variance is available.[17] The hypothesis testing calculations were performed using the average SLA values for the groups, and the results were transformed to fold change by using the equivalence of $|\log(SA_i) - \log(SA_j)| = d$ and $SA_i/SA_j = 10^{|d|}$, where SA_i denotes the average SA in group i, d *is* the size of the effect difference between the SLA in the two groups, and $F_{ij} = SA_i/SA_j$ is the fold change between the groups. Thus, a difference of d in the SLA averages corresponded to a fold change of $f = 10^{|d|}$.

Further, the total variance at a spot was deconvolved into components of sample preparation, dye-to-dye differences, or gel-to-gel variations by fitting a mixed-effects model.[18] The

$$y_{ijk} = \mu + \alpha_i + a_j + b_k + e_{ijk} \quad (1)$$

model was fit, where y_{ijk} denotes the standardized log abundance response at a fixed spot, μ denotes the overall mean; α_i are the fixed dye effects for $i = 1, 2$; a_j are the random gel effects for $j = 1, ..., 12$; b_k are the random sample effects for $k = 1, ..., 8$; and e_{kij} is a random error term. The sample and gel effects were modeled as random, since both the samples and gels represented random samples from larger populations, while the dye effect was assumed to be fixed to reflect the fixed nature of the dyes. Further normality assumptions included $a_j \approx N(0,\sigma_g^2)$, $b_k \approx N(0,\sigma_s^2)$, $e_{ijk} \approx N(0,\sigma_e^2)$, distributed independently of each other. The total variance σ_y^2 at a spot was thus decomposed into three random components, the gel variance σ_g^2, the sample variance σ_s^2, and the error variance σ_e^2: $\sigma_y^2 = \sigma_g^2 + \sigma_s^2 + \sigma_e^2$. The existence of a fixed dye effect was assessed with a usual analysis of variance (ANOVA) test, and the variance components were quantified by comparing their estimates.

The effect of statistical normalizationsof the SLA on the spot-wise standard deviations and on the number of differentially expressed spots was investigated. Briefly, the SLA values obtained from DeCyder were additionally normalized by statistical methods that corrected for potential dye biases within gels and range differences among the gels.[19] Following data analysis, the proteins of interest with variable expression levels across the gels were robotically excised and identified by mass spectrometry (MS) (Proteomic Research Services, Ann Arbor, MI) as reported[14] (also found in Mahnke, R. C.; Corzett, T. H.; McCutchen-Maloney, S. L.; Chromy, B. A. *J. Proteome Res.*, in press).

Results and Discussion

Spot Matching. Manual inspection verified that the master gel, which contained the highest number of spots, identified by the batch processor contained no visible defects such as bubbles, dust, or precipitated dye. Following the manufacturer's recommended protocol, manual landmarks were used to assist in the spot matching across the gels. As a single individual performed all of the landmarking, the effect of this manual process on the results was not modeled. Spots of interest identified through the analyses to be differential were manually verified to have the three-dimensional profile characteristics of a protein spot. By the manual verification, spots with volume measurements close to the background and dust particles with no defined shape were eliminated. We stress that the statistics were calculated using all spots, and manual verification was only used for determining spots for identification. The total number of spots detected on the master gel was 2511. One hundred and sixty-nine (6.8%) of those spots were detected and matched on all 12 gels. Three hundred and thirty-six (13.4%) of the spots were matched on at least 11 gels, 549 (21.9%) matched on 10 or more gels, 797 (31.7%) matched on at least 9 gels, and 1055 (42.1%) matched on at least 8 gels. Subsequent analyses were restricted to the 1055 well-matched spots that matched on at least 8 gels.

Previous matching results with DIGE are only available with *E. coli* and range from 52% of the spots matched on 12 of 12 gels[10] to 67% matched on 8 of 8 gels.[23] The results here reflect a decreased spot matching accuracy with human serum sample aliquots. By using gels with pH 3−10 strips, we covered in this study a wide range of proteins, 20−220 kDa, for a global

overview of the human plasma proteome. Focusing on a narrower isoelectric point and molecular weight range is expected to result in better matching.

Patterns of Differential Expression and Relation to Top-6 Depletion. After making all pairwise sample-to-sample comparisons using DeCyder, 144 unique spots were found to be differential, of which 37 were verified to be valid protein spots following manual inspection. The rest of the spots were determined to be artifacts, as they did not exhibit the required protein spot characteristics. Running an artifact-exclusion filter on the original spots before performing the statistical analyses would have reduced the number of differentially expressed spots. However, to minimize the number of potentially missing valid protein spots, our strategy was to include all spots in the initial analysis and eliminate the artifacts manually. A similar comparison between the samples labeled with Cy3 and Cy5 revealed 168 differentially expressed spots, only three of which passed manual verification. A direct pairwise comparison between the 12 gels resulted in 252 spots showing gel-to-gel variation, of which 28 passed manual verification. The three types of variation resulted in 53 unique spots (2.1% of the 2511 spots on the master gel) showing differential levels of protein. Of the 53, 29 had gel-to-gel variation and 3 had only dye-to-dye variation, leaving 21 spots with unexplained sample-to-sample variation. In a true biological experiment, 21 of the 2511 spots (0.8%) could have been deemed as showing differential expression. Of these 21 spots, the 10 that were verified on the preparative gel were chosen for identification by mass spectrometry.

Expression pattern clustering in EDA found three distinct patterns within the 21 differentially expressed spots. The first pattern, observed in five spots (1181-Albumin, 1234-Albumin, 1244, 1250, and 1255), showed a difference between sample A, the first sample processed through the top-6 depletion, and the other seven samples. Spot number 1250, for example, had a 1.7-fold decrease between sample A and the others, with an observed T-test p-value of 5.1×10^{-12}. These five spots were grouped around a region of albumin migration. The two spots that were identified by MS were confirmed to be albumin. The second pattern, observed in 10 spots (646, 684-Plasminogen, 686, 874, 878, 961, 962-Transferrin, 968-Transferrin, 969-Transferrin, and 979-Transferrin), showed a decreased level of protein for samples A, D, and G (processed 1st, 4th, and 7th through the top-6 depletion, respectively) relative to the other sample aliquots. These spots clustered in a region of transferrin migration.

The third pattern, similar to the second, showed a decreased level of protein in samples A, D, and G. However, the five spots in this group (1783, 1999, 2018, 2187, 2364-Proapolipoprotein) were randomly distributed across the gels. One spot, 2138, identified as Amyloid P component, showed an increase in protein for samples A, D, and G relative to the other samples. Four of these six spots (1783, 1999, 2018, 2187, 2364-Proapolipoprotein, and 2138-Amyloid P component) that were not close to the regions of albumin or transferrin on the gels could not be picked for identification because they were not matched on the pick gel. However, as their patterns were similar to those of transferrin, this suggested that their differences were associated with variability in the depletion of the top-6 proteins in the top-6 depletion step. Two of the proteins that were identified, 2364-Proapolipoprotein and 2138-Amyloid P component, have been shown to be associated with transferrin.[20,21]

While previous studies found that the Agilent Multiple Affinity Removal System selectively depleted the top-6 high-abundance proteins reproducibly,[22] here differential levels of albumin and transferrin were observed between technical replicates of the same sample (as large as 3.4-fold for spot 1234), indicating large variability associated with the top-6 depletion step during sample preparation. Closer examination of the depletion procedure revealed that the decreased levels of albumin in sample A corresponded to the first sample run over the Agilent Multiple Affinity Removal System. Our standard protocol for the Agilent Multiple Affinity Removal System calls for an equilibration run, in which a blank sample consisting of Agilent buffer A is injected to verify that the column has been regenerated from the previous runs, that no residual protein is still bound to the column, and to establish a baseline for the experimental samples that follow. The decreased levels of albumin present in the flow-through fraction suggested that the equilibration run was creating a "super sensitivity" to albumin or the ability to remove additional albumin, due to the additional washing of the IgG antibodies. Removing the first equilibration run before processing the samples is expected to decrease the variation. However, the addition of the equilibration run has several additional benefits, including the increased depletion of albumin. A better solution would be to add an equilibration run between each sample, similar to the blank injections used by Martosella and colleagues to test for protein carryover.[22] In addition to the ability to verify that no residual protein is still bound to the column and to establish a baseline, the super sensitivity to albumin will allow for increased protein depletion. Here, the small variation observed within the samples other than sample A is attributed to the protein depletion step, and adding the equilibration run would likely reduce this variable.

Prediction Intervals. The distribution of the SLA indicated consistent data across the gels and dyes. In terms of expression ratios, the 2.5th percentiles over the 12 gels and two dyes were in the [−1.23, −3.55] interval when using all the spots, and in [−1.21, −1.42] when considering only the well-matched spots, defined as the spots matched on at least eight gels. The corresponding 97.5th percentiles were [1.54, 8.25] and [1.24, 1.51], respectively. Considering all spots, the 95% prediction interval for the expression ratios on a future gel was [−2.55, 6.87]. When using only the well-matched spots matched on at least eight gels, the 95% prediction interval decreased to [−1.84, 1.90]. Thus, under these assumptions, an expression ratio on a future human plasma gel will have to be increased or decreased at least 1.9-fold in intensity to signify a true difference in plasma concentration at the 95% confidence level. This requirement is higher than the best-case scenario 1.2-fold change reported with mouse liver homogenates[9] and reflects the added complexity of the human plasma samples.

Spot-Wise Variation. The coefficient of variation of the SA was calculated for the well-matched spots (Figure 1a). At each spot, the average and the SD were taken over the 24 values that corresponded to the two SA measurements (corresponding to Cy3/Cy2 and Cy5/Cy2) over the 12 gels. Results considering the two sets of 12 values corresponding to the two dye combinations separately were similar (data not shown). The spots were then grouped into three sets according to the matching results: those matched on all 12 gels, those matched on 10 or 11 gels, and those matched on 8 or 9 gels. As the three subsets had varying sizes, the corresponding CV distributions (Figure 1a) were compared in terms of percentiles. The three sets showed similar characteristics for up to the 50th percen-

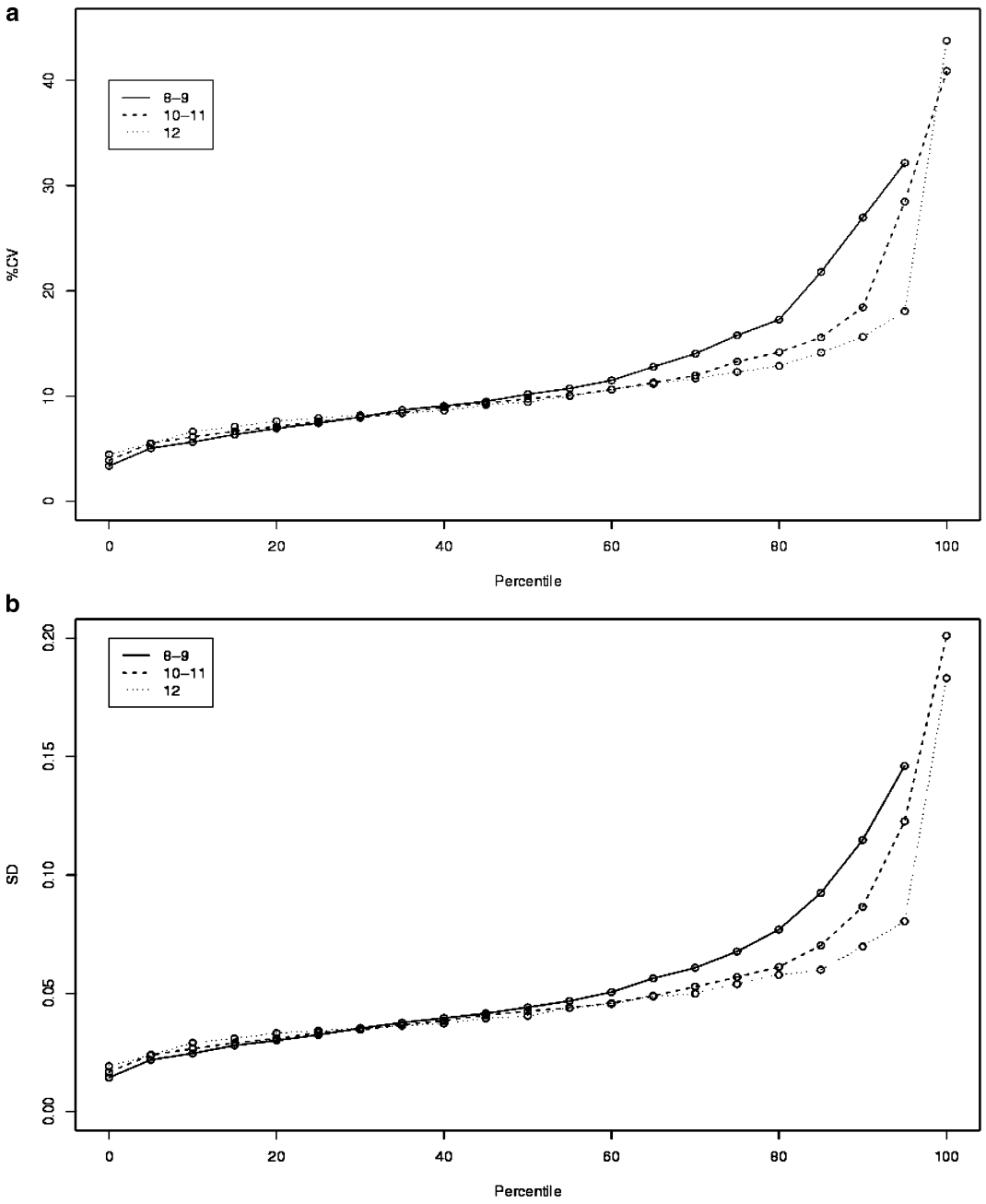


Figure 1. (a) The spot-wise %CV values of the SA, for the 1055 spots matched on at least 8 gels. The three curves correspond to different subsets of the spots: matched on 8 or 9 gels (506 spots), 10 or 11 gels (380 spots), or on all 12 gels (169 spots). The maximum value (107.56) of the 8−9 set was omitted for clarity. (b) The spot-wise SD values of the SLA, for the 1055 spots matched on at least 8 gels, and selected percentiles of the spot-wise SD distributions of the SLA. The three curves correspond to different subsets of the spots: matched on 8 or 9 gels (506 spots), 10 or 11 gels (380 spots), or on all 12 gels (169 spots). The maximum value (0.418) of the 8−9 set was omitted from the graph for clarity.

tiles, with all CV values below 10%. At the higher percentiles, the subset based on the spots matched on all 12 gels had slightly, but consistently, lower CVs than the other two subsets. For all three subsets, less than 20% of the spots had CVs exceeding 15%. The highest CV for the spots matched on all 12 gels was 43.8%. For the spots matched on 10 or 11 gels, the maximum CV was 40.9%. For the spots matched on 8 or 9 gels, one potentially mismatched spot resulted in the maximum CV of 107.6%.

Spot-wise SD values of the SLA were determined similarly for the well-matched spots (Figure 1b). For all three subsets, 50% of the spots had SD less than 0.044. For the higher percentiles, the subset matched on all 12 gels had slightly lower SDs than the corresponding values for the other two subsets. The maximum SD was 0.183 for the subset matched on all 12 gels, 0.201 for the spots matched on 10 or 11 gels, and 0.418 for the spots matched on 8 or 9 gels. The spatial distribution of the spot-wise SDs (Figure 2a) differentiated regions with low (red to orange) and high (blue to yellow) variation, which correlated with the quality of the spot matching (Figure 2b). Spots with smaller SD values tended to be in regions of the gels with better matching characteristics. It is also apparent (Figure 2b) that the spots with poor matching (matched on fewer than eight gels) were mostly concentrated around the

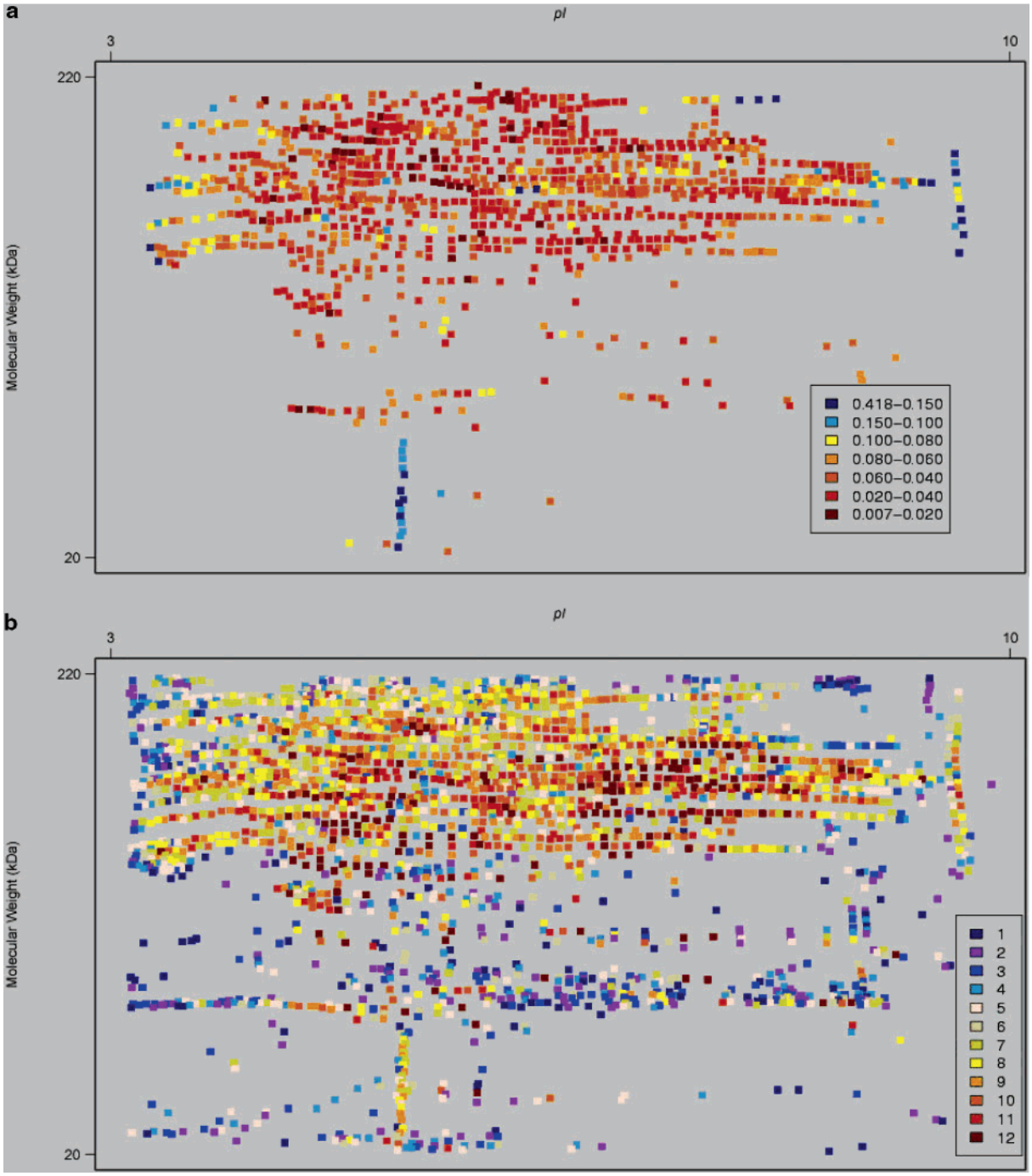


Figure 2. Spatial distribution of SD values and spot matching. (a) Shown for 1055 spots matched on at least 8 gels. The SD values were binned into the seven intervals shown in the legend. (b) Shown for all 2511 spots. The matching varies from 1 (spots found only on the master gel) to 12 (spots matched on all 12 gels), and spots are color-coded according to the legend. The *Y*-axis represents molecular weight (20−220 kDa), and the *X*-axis represents p*I* range (3−10 nonlinear).

edges of the gel. In addition, many of the spots that showed low SD values but high matching scores, most notably the vertical trails at the right and bottom of the gels, were determined to be artifacts and manually excluded during the protein identification process. Summary statistics of the spot-wise SDs of the SLA confirmed that the better-matched spots exhibited less variation over the gels (Figure 1b).

The SD estimates for the SLA agreed closely with a recent 2-D DIGE study using six same-same gels with *E. caratova* bacterial cells, where the 1st, 25th, 50th, 75th, and 100th percentiles of the spot-wise SD values were 0.011, 0.027, 0.043, 0.067, and 0.223, respectively.[13]

Sample Size and Power. The SD estimates for the SLA were used in statistical sample size and power calculations for future 1-factor ANOVA experiments involving human plasma. Table 2 presents the minimum number of replicates required at each level of the factor to detect an *f*-fold change in the average ratio between two treatments with $\alpha = 0.01$ and power = 0.80, for varying levels *k* of the factor, based on select standard deviation estimates from Figure 1b. For example, using five groups to

Table 2. The Minimum Number of Replicates *r* Required to Detect an *f*-Fold Average Ratio Change between Two Samples in a 1-Factor ANOVA Experiment with *k* Levels of the Factor, and the Select Standard Deviation Estimates from Figure 1b Specified under the Spot Set and Percentile Headings

		k											
		50th percentile						75th percentile					
spot set	fold change *f*	2	3	4	5	6	7	2	3	4	5	6	7
12	1.2	8	9	10	11	11	12	13	15	16	17	18	19
	1.5	4	4	4	4	4	4	5	5	5	5	5	5
	2.0	3	3	3	3	3	3	3	3	3	3	3	3
10−11	1.2	9	10	11	11	12	13	14	16	18	19	20	21
	1.5	4	4	4	4	4	4	5	5	5	5	5	5
	2.0	3	3	3	3	3	3	3	3	3	3	3	3
8−9	1.2	10	11	12	13	13	13	19	22	24	26	28	29
	1.5	4	4	4	4	4	4	6	6	7	7	7	7
	2.0	3	3	3	3	3	3	4	4	4	4	4	4

Table 3. The Minimum Fold Change *f* Detectable under the Same Assumptions as in Table 2 with *k* Levels of the Factor and *r* Replicates

		k											
		50th percentile						75th percentile					
spot set	replicates *r*	2	3	4	5	6	7	2	3	4	5	6	7
12	3	1.6	1.6	1.6	1.6	1.6	1.6	1.8	1.8	1.8	1.8	1.8	1.9
	4	1.4	1.4	1.4	1.5	1.5	1.5	1.6	1.6	1.6	1.6	1.6	1.6
	5	1.3	1.4	1.4	1.4	1.4	1.4	1.5	1.5	1.5	1.5	1.5	1.5
10−11	3	1.6	1.6	1.6	1.6	1.6	1.6	1.9	1.9	1.9	1.9	1.9	1.9
	4	1.4	1.5	1.5	1.5	1.5	1.5	1.6	1.6	1.6	1.7	1.7	1.7
	5	1.4	1.4	1.4	1.4	1.4	1.4	1.5	1.5	1.5	1.5	1.6	1.6
8−9	3	1.7	1.7	1.7	1.7	1.7	1.7	2.1	2.1	2.1	2.1	2.1	2.1
	4	1.5	1.5	1.5	1.5	1.5	1.5	1.7	1.8	1.8	1.8	1.8	1.8
	5	1.4	1.4	1.4	1.4	1.4	1.4	1.6	1.6	1.6	1.7	1.7	1.7

test the effects of five different samples corresponds to $k = 5$. The median and the 75th percentile of the SDs were selected as example estimates of the error SD in future experiments. To detect a 2-fold change in the protein expression ratio between two groups with a 1% significance level and 80% power, at least four replicates per group are required, when using 0.068 (the 75th percentile of the spot-wise SD distribution for the spots matched on 8 or 9 gels) as the estimate of the SLA error SD and the protocol reported here. Under the same conditions, six replicates are required to detect a 1.5-fold change, and 19 replicates to detect a 1.2-fold change. These values are comparable to recent minimum sample size results[13] based on bacterial cells (four replicates for detecting a 2-fold change, seven for a 1.5-fold change, and 18 for a 1.25-fold change), where the error variance was estimated by the average noise seen in 75% of the spots, and the same 1% significance level and 80% power parameters were used.

Complementary to Table 2, Table 3 indicates the fold changes that can be detected given the indicated number of treatments and replicates, with the same 99% confidence, 80% power, and standard deviation estimates as in Figure 1b. The most conservative values correspond to the 8 or 9 spot set, but with higher-quality data (for example, the 12 set, in which a protein spot is matched on all 12 gels), smaller fold changes become significant.

Variance Decomposition Using Mixed-Effects Statistical Models. The model in eq 1 was fit to the 1055 well-matched spots matched on at least eight gels. The resulting *p*-values for the dye effect were adjusted for multiple comparisons with the FDR method.[24] One hundred ninety-eight spots had adjusted *p*-values ≤ 0.05. From those, 40 spots had at least a 1.5-fold difference between the average responses under the Cy3 and Cy5 dyes. The statistical model considered simultaneously all sources of variation and dramatically reduced the estimated number of spots with a dye effect from the number obtained with the simpler method available in DeCyder (40 vs 168).

Table 4. The Frequency Distribution of the Variance Component Estimates from Eq 1 for the 1055 Spots Matched on at Least 8 Gels[a]

% contribution to total variance	sample		gel		random error	
	(a)	(b)	(a)	(b)	(a)	(b)
0−10	36.11	36.11	7.20	7.20	49.85	49.85
10−20	7.96	44.08	8.91	16.11	46.07	95.92
20−30	7.68	51.75	10.05	26.16	4.08	100.00
30−40	8.63	60.38	10.9	37.16		
40−50	9.48	69.86	9.86	47.01		
50−60	8.06	77.91	7.20	54.22		
60−70	7.96	85.88	7.49	61.71		
70−80	6.35	92.22	4.64	66.35		
80−90	5.97	98.20	4.17	70.52		
90−100	1.80	100.00	29.48	100.00		

[a] The components of sample preparation, gel-to-gel differences, and random error are shown separately as (a) the percentage of spots and (b) the cumulative percentage of spots with contribution to the total variance indicated in the first column.

The random variance component estimates from eq 1 indicated that, for most spots, the largest component of variation was due to gel-to-gel differences (Table 4). For close to 30% of the spots, the gel variance component was over 90% of the total variance. For only 47% of the spots, the gel variance component was less than 50% of the total variation. In contrast, for close to 70% of the spots, the variation due to sample preparation contributed less than 50% to the overall variation, and for less than 2% of the spots, the sample preparation component was over 90% of the total variation. The unexplained error component was the smallest, contributing less than 30% to the total variation at each spot. The results are visually displayed for the spots matched on all 12 gels (Figure 3). For the majority of spots with large gel-to-gel variation, extreme values on one gel compared to all the other gels were responsible. For example, spot 1116, with the highest variance (Figure 3), had its two observations on gel 4 twice as large as its values on any of the other gels. Such discrepancies could be due to potential spot mismatches or to other differences among the gels. The spots with largest sample variation (Figure 3) corresponded to the albumin- and transferrin-related regions identified previously through clustering and MS identification.

Statistical Normalization. The spot-wise SDs of the normalized[19] SLA values for the spots matched on at least eight gels ranged from 0.016 to 0.417, with a median of 0.041. To test the effect of the normalizations on the number of spots showing differential expression among the samples, a 1-factor fixed-effect ANOVA model with 8 levels corresponding to samples A−H was fit at all spots. Table 5 summarizes the number of spots that satisfied the following criteria: at least 1.5-fold change between any two sample means and ANOVA *p*-value ≤ 0.05. The columns differentiate the type of data (the SLA or the additionally normalized SLA) and the subset of spots (all spots or only spots matched on at least eight gels) used in the analysis. The rows indicate whether original or FDR-adjusted[24] *p*-values were used.

The additional normalizations resulted in only a slight decrease in the median spot-wise standard deviation of the

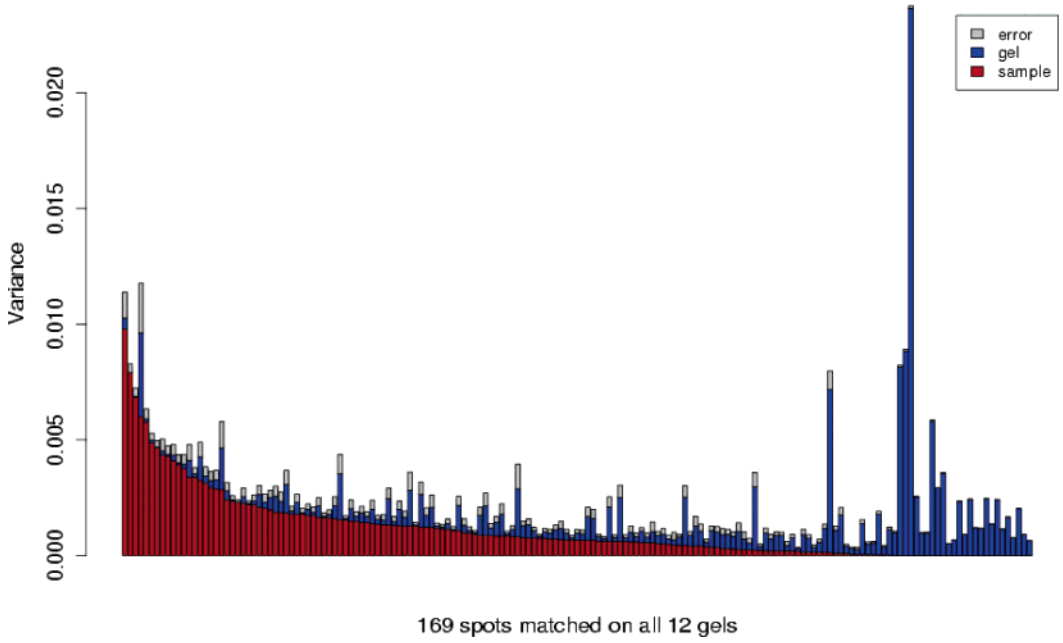


Figure 3. The variance component estimates from eq 1 for the 169 spots matched on all 12 gels, ordered by the estimated sample variance.

Table 5. The Number of Spots with a Greater Than 1.5-Fold Change between Any Two Samples and with an ANOVA Sample Effect *p*-Value Less than 0.05, as a Function of the Variable (SLA or Normalized SLA), Subset of the Spots (All or Well-Matched), and the Type of *p*-Value (Original or Adjusted) Used

	SLA		statistically normalized SLA	
	all spots $n = 2511$	well-matched spots $n = 1055$	all spots $n = 2511$	well-matched spots $n = 1055$
p-value	117 (4.66%)	53 (5.02%)	74 (2.95%)	29 (2.75%)
FDR adjusted *p*-value	33 (1.31%)	36 (3.41%)	16 (0.63%)	17 (1.61%)

well-matched spots (0.041 vs 0.043), suggesting that the original SLA were adequately normalized across the gels. Although the extra normalization did not decrease the overall variation, it reduced dramatically the number of spots with an estimated sample effect (Table 5). When using the well-matched spots and the adjusted *p*-value, the reduction was over 50% (17 vs 36). The 17 spots found to be differential across the samples were a fraction of the 144 spots identified to have a sample effect with DeCyder. Interestingly, the 17 spots, found to be differentially expressed, comprised the bulk of the 21 proteins identified through extensive manual inspection after the initial analysis with DeCyder.

Conclusions

A detailed analysis of a 2-D DIGE experiment using one human plasma sample prepared eight times by high-abundance protein depletion and analyzed in triplicate on 12 gels found that, considering the subset of spots matched on at least 75% of the gels, over 50% of the spots had less than 10% CV for the SA, and less than 20% of the spots had larger than 15% CV of the SA. Spot-wise SD values of the SLA, and subsequent power and sample size calculations, were in close agreement with recent results based on mouse brain, liver, and heart homogenates and bacterial cell samples.[13] The reported SD values can be used to estimate the standard deviation of the error in power and sample size calculations for more complex biological experiments with a similar protocol. The 75th percentile of the SD values of the subset of spots matched on 8−9 gels provides a conservative noise SD estimate for such calculations. However, more stringent results can be obtained for the well-matched spots that are more consistently matched on the gels. Results with the conservative estimate indicate that fold changes greater than 2 can be detected with a manageable number of replicates in simple ANOVA experiments with human plasma.

From the present study, assuming that the error variance in future biological studies is similar to the technical variation found here, the reported technical replicates can be used to determine the number of biological replicates needed for proteomic characterization of human plasma. For example, if one were comparing two populations, healthy versus diseased, with the protocol reported here, using the most conservative case, four biological replicates would be required from both populations to detect a 2-fold change. Including additional technical replicates from the biological replicates would provide additional information for estimating the within-technical-replicates variation for that experiment. Fewer biological replicates, or lower fold-changes for a given number of replicates, would be needed for the less conservative case where the 12 subset is applicable.

Statistical mixed-effects modeling quantified the relative contribution of the sample preparation, gel differences, and random error to the total variance. Gel-to-gel differences were found to comprise the largest component, followed by the sample preparation. Future improvements in gel quality and innovations in spot detection and matching have the potential to further reduce the gel-to-gel variation. The top-6 high-abundance protein depletion was identified as the reason for the large variance component associated with the sample preparation. As most differentially expressed proteins identified in this study were related to albumin and transferrin, caution should be exercised in interpreting results of future biological

experiments performed with a previously reported protocol for top-6 most abundant proteins.[14] A modified top-6 depletion protocol is suggested that includes adding an equilibration liquid chromatography step between each plasma sample processed through the Agilent Multiple Affinity Removal System. This approach minimizes sample-to-sample experimental variation and provides an improved method to analyze human plasma by 2-D DIGE and other proteomic platforms, in particular with regard to population proteomics in which multiple plasma samples are processed simultaneously for comparative differential expression analysis.

Acknowledgment. This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, with support from the Department of Homeland Security (Biological Countermeasures Program) and Laboratory Directed Research and Development funding UCRL-JRNL-219771.

References

(1) Jacobs, J. M.; Adkins, J. N.; Qian, W. J.; Liu, T.; Shen, Y.; Camp, D. G.; Smith, R. D. Utilizing human blood plasma for proteomic biomarker discovery. *J. Proteome. Res.* **2005**, *4*, 1073−1085.

(2) Nedelkov, D. Population proteomics: addressing protein diversity in humans. *Expert Rev. Proteomics* **2005**, *2*, 315−324.

(3) Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B. B.; Simpson, R. J.; Eddes, J. S.; Kapp, E. A.; Moritz, R. L.; Chan, D. W.; Rai, A. J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W. S.; Hefta, S. A.; Meyer, H.; Paik, Y. K.; Yoo, J. S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C. Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; Andrews, P.; Tammen, H.; Speicher, D. W.; Hanash, S. M. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly available database. *Proteomics* **2005**, *5*, 3226−3245.

(4) Omenn, G. S.; Paik, Y. K.; Speicher, D. The HUPO Plasma Proteome Project: a report from the Munich congress. *Proteomics* **2006**, *6*, 9−11.

(5) Omenn, G. S. Exploring the human plasma proteome. *Proteomics* **2005**, *5*, 3223−3225.

(6) O'Farrell, P. H. High-resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **1975**, *250*, 4007−4021.

(7) Lilley, K. S.; Razzaq, A.; Dupree, P. Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Curr. Opin. Chem. Biol.* **2002**, *6*, 46−50.

(8) Ünlü, M.; Morgan, M. E.; Minden, J. S. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **1997**, *18*, 2071−2077.

(9) Tonge, R.; Shaw, J.; Middleton, B.; Rowlinson, R.; Rowlinson, R.; Rayner, S.; Young, J.; Pognan, F.; Hawkins, E.; Currie, I.; Davison, M. Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics* **2001**, *1*, 377−396.

(10) Alban, A.; David, S. O.; Bjorkesten, L.; Andersson, C.; Sloge, E.; Lewis, S.; Currie, I. A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics* **2003**, *3*, 36−44.

(11) Lilley, K. S.; Friedman, D. B. All about DIGE: quantification technology for differential-display 2D-gel proteomics. *Expert Rev. Proteomics* **2004**, *1*, 401−409.

(12) Marouga, R.; David, S.; Hawkins, E. The development of the DIGE system: 2D fluorescence difference gel analysis technology. *Anal. Bioanal. Chem.* **2005**, *382*, 669−678.

(13) Karp, N. A.; Lilley, K. S. Maximising sensitivity for detecting changes in protein expression: experimental design using minimal CyDyes. *Proteomics* **2005**, *5*, 3105−3115.

(14) Chromy, B. A.; Gonzales, A. D.; Perkins, J.; Choi, M. W.; Corzett, M. H.; Chang, B. C.; Corzett, C. H.; McCutchen-Maloney, S. L. Proteomic analysis of human serum by two-dimensional differential gel electrophoresis after depletion of high-abundant proteins. *J. Proteome Res.* **2004**, *3*, 1120−1127.

(15) Echan, L. A.; Tang, H. Y.; Ali-Khan, N.; Lee, K.; Speicher, D. W. Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. *Proteomics* **2005**, *5*, 3292−3303.

(16) Venables, W. N.; Smith, D. M.; and the R Development Core Team. *An Introduction to R: Notes on R, A Programming Environment for Data Analysis and Graphics, v.2.0.1*; Network Theory Ltd.: Bristol, U.K., 2004.

(17) Kuehl, R. O. *Design of Experiments: Statistical Principles of Research Design and Analysis*, 2nd ed.; Duxbury Press: Pacific Grove, CA, 2000.

(18) Pinheiro, J. C.; Bates, D. M. *Mixed-Effects Models in S and S-PLUS*; Springer Statistics and Computing: New York, 2000.

(19) Fodor, I. K.; Nelson, D. O.; Alegria-Hartman, M.; Robbins, K., Langlois R. G.; Turteltaub, K. W.; Corzett, T. H.; McCutchen-Maloney, S. L. Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyder. *Bioinformatics* **2005**, *21*, 3733−3740.

(20) Kunitake, S. T.; Jarvis, M. R.; Hamilton, R. L.; Kane, J. P. Binding of transition metals by apolipoprotein A-I-containing plasma lipoproteins: inhibition of oxidation of low-density lipoproteins. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 6993−6997.

(21) Hamazaki, H. Structure and significance of N-linked sugar unit of human serum amyloid P component. *Biochim. Biophys. Acta* **1990**, *1037*, 435−438.

(22) Martosella, J.; Zolotarjova, N.; Liu, H.; Nicol, G.; Boyes, B. E. Reversed-phase high-performance liquid chromatographic prefractionation of immunodepleted human serum proteins to enhance mass spectrometry identification of lower-abundant proteins. *J. Proteome Res.* **2005**, *4*, 1522−1537.

(23) Yan, J. X.; Devenish, A. T.; Wait, R.; Stone, T.; Lewis, S.; Fowler, S. Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*. *Proteomics* **2002**, *2*, 1682−1698.

(24) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B.* **1995**, *57*, 289−300.

PR060100P

UCRL-ID-151868

State-Based Automata Descriptions of Intracellular Protein Kinetics and Gene Regulation

J. R. Kercher

January 1, 2003

U.S. Department of Energy
Lawrence Livermore National Laboratory

DISCLAIMER

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

State-based Automata Descriptions of Intracellular Protein Kinetics and Gene Regulation

J.R. Kercher
Environmental Sciences Division, L-396
Energy and Environment Directorate
Lawrence Livermore National Laboratory, P.O. Box 808
Livermore, California 94550

INTRODUCTION

States. Turteltaub et al. (2002) use the words "components of pathogenicity" and "stages" to describe discrete activities of infection and discrete time periods of infection in which certain specific events occur. In other contexts these same concepts might be labeled *states.* In this report I consider a formal concept of *state* to capture the idea of the cell or genetic subsystem being in a discrete, and more or less unique or specific, activity with a distinct set of constituents. For example, Cornelis (1998) reviewed the *Yersinia* literature and reported that when *Yersinia* spp. transits from a temperature and calcium regime found in the flea blood to that of human tissue, it undergoes a change of state; a new set of proteins appear (Yops). We suggest that the state-based approach is useful for describing, predicting, and understanding the significance to cell functioning of discrete changes in the gene complex and resulting protein complex within the cell.

Information processing by the gene-protein complex. While information is stored genetically, a complicated apparatus exists that controls which information is accessed (to make proteins) by regulated genes. This apparatus is under at least partial control by proteins produced by other genes. At least to some degree, these proteins, which provide feedback to control access to the "genetic source code", may respond to environmental conditions or substrates either inside the cell or at the cell surface. These relationships seem to suggest that we may regard the cell system (or at least the gene-protein complex) as an information processor, which senses its environment and adjusts its activities (state) in accord with certain rules imposed biochemically.

Furthermore, if the gene-protein complex is an information processor to some degree, then it may be useful to explore information-processing theory to help understand the gene-protein relationships of our model systems. Specifically, it may be possible that information-processing theory can inform the interpretation or analysis of the potentially huge amounts data to be gathered in experiments of gene expression and protein production. Ultimately, we may be able to exploit information-processing concepts to aid experiment selection.

Automata in information processing theory. In formal computation theory, there are four classes of mathematically rigorous machines or automata that can process information. In order of increasing computational power, these machine classes are:

a) finite automata (FA), also known as finite state automata (FSA)

b) push down automata (PDA). A PDA is an FA with an infinite push down stack.

c) linear bounded automata (LBA)

d) Turing machines (TM). A TM is the same as an LBA except that the tape is infinite.

There is a rigorous mathematical definition for each machine, and these machines are well studied and well described (e.g., Hopcroft and Ullman 1979, Taylor 1998). Furthermore each machine class is associated with a specific

language class. As the machines progress in computing power, the associated languages grow in complexity. Anything that can be computed by FA's can be computed by PDA's; anything that can be computed by PDA's can be computed by LBA's; anything that can be computed by LBA's can be computed by TM's. Conversely, TM's can compute (or recognize) functions that LBA's cannot; LBA's can recognize sequences of symbols that PDA's cannot, and PDA's can recognize sequences of symbols that FA's cannot.

We suggest that two of these machines, finite state automata (FA) and linear-bounded automata (LBA), are candidate formalisms for describing the internal dynamics of cellular functioning.

As component of larger model. In our view, gene expression models are one of the component submodels of the larger model of cellular functioning or intercellular signaling, which is our over-arching goal. For those cases in which the relaxation time of gene expression system is short compared the characteristic times of the other processes in the larger model, describing the results of gene expression as discrete states may be a desirable alternative to descriptions based on ordinary differential equations (ODEs). In this instance, a state-based approach helps avoid stiff ODEs.

Previous work. Other authors have proposed state-based approaches or finite automata as methods for describing gene expression. Feitelson and Treinin (2002) have proposed that the cell be regarded as an FA, whose states are defined by the protein constituents and whose control mechanism is the genetic network. Somogyi et al. (2001) have developed algorithms for inferring genetic networks from gene expression data. Their method relies on Boolean states (on/off) as the fundamental variable. Boolean networks are a common form of encoding gene regulation (Liang et al. 1998, Thieffry and Thomas 1998, Hatzimanikatis and Lee 1999). Thomas and D'Ari (1990) carefully consider the time constants of the various processes in a gene regulation network. Their methodology follows the time development of meta-stable states following on the change of some external variable. The centerpiece of their approach is to make a distinction between the state of the gene and the state of the product (proteins). This is similar to the distinctions made below regarding writing on the "tape", the "calculation states", and the final states.

Defining automata. Informally, a finite state automaton (FA) consists of three pieces: (1) a one-way read-only tape, (2) a finite set of states, and (3) a

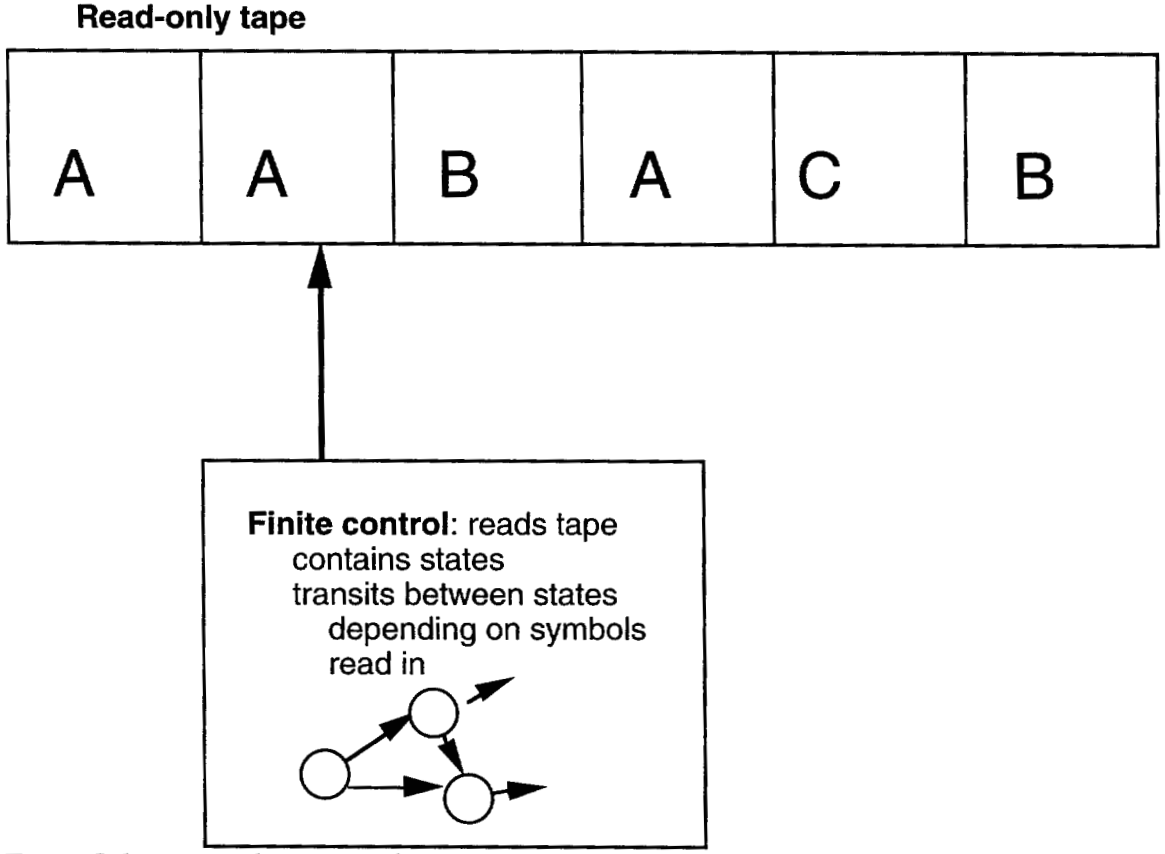


Fig. 1 Schematic diagram of a Finite State Automaton.

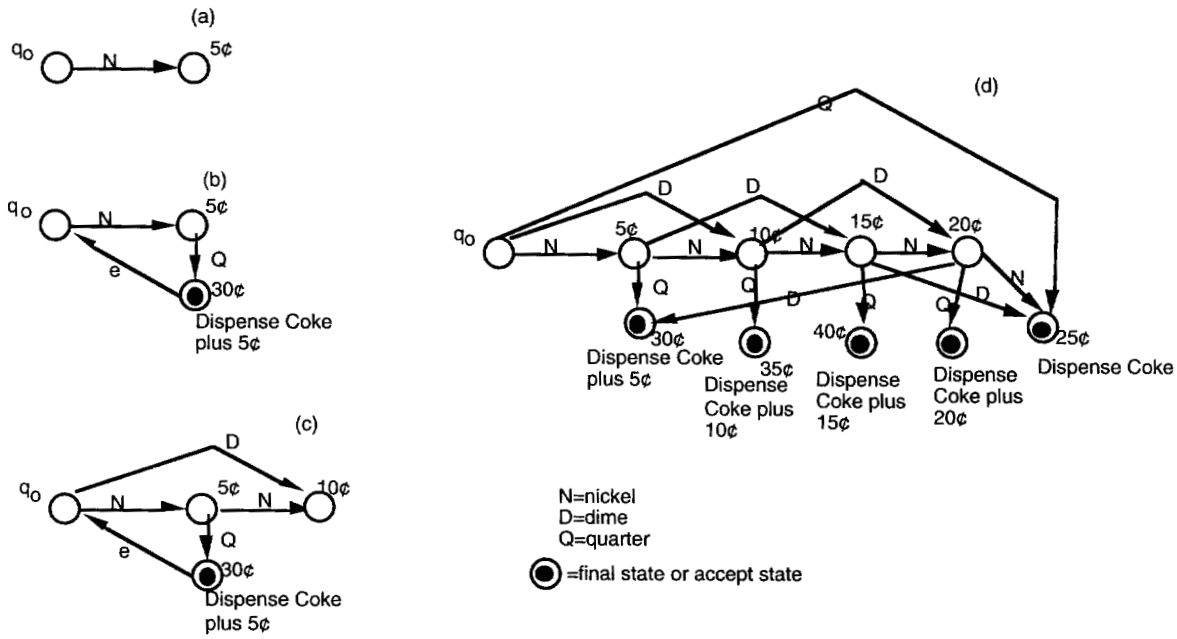


Fig. 2. State diagrams for Coke machine for which 25¢ dispenses one Coke. (a) move to 5¢ state, (b) adding a quarter, (c) two transfers to 10¢ state, (d) total state-transition diagram.

control system that reads the tape and changes the state of the machine depending on a pre-determined set of rules, which are based on the current state and the symbol read from the tape. This is shown schematically in Fig. 1. Formally, the mathematical description of a finite state automaton consists of Q, the set of states; Σ the input alphabet on the tape; δ a transition function that describes the transitions between the states; and the set of final states *F* that determine whether the machine recognizes the input string on the tape. One of the states is a unique start state q_0. A linear-bounded automata (LBA) differs from an FA in that (1) the head, which reads the tape, can move in both directions rather than in just one, (2) the head can both read and write on the tape, and (3) the transition rules specify the moves between the states, the symbols to be written, and the moves of the head.

The significance of computing machines (FA's and LBA's) for the study of gene expression and protein production. In both types of automata, the notion of state is central to their structure and operation. *The state of the gene-protein complex we take to be the proteins produced and the concentration of their substrates in the cell.* Thus these state-based computational schemes provide a natural description of the gene-protein complex because each can be built around the quantities (gene expression and protein occurrence) that are directly observable by experiments. Furthermore, these systems (FA's and LBA's) and the types of languages, which they accept, are well studied. As stated above, the known properties of the automata classes and their languages might be exploitable in designing experiments.

A FAMILIAR EXAMPLE OF A FINITE STATE AUTOMATON

The 25¢ Coke machine finite state automaton. An example of a finite state automaton is the Coke machine in the hall in B 361. (In this case, instead of a read-only tape, the input is the coins put into the slot.) Let's say the Coke machine dispenses a Coke at 25¢ (ca 1963) to keep the description simple. The machine starts in the start state q_0 and suppose someone deposits a nickel; then the machine transitions to the "5¢-state", which means "Someone has deposited a total of 5¢" (Fig. 2a). Now if the machine is in the 5¢-state and the person deposits a quarter (Fig. 2b), then machine transfers to the 30¢-state (a final state or accept state) and dispenses one Coke and 5¢ in change. If instead, the person had deposited another nickel when the machine was in the 5¢-state, the machine would have gone to the 10¢-state, Fig. 2c. Alternatively, when the machine was in the initial state, a deposit of one dime would have moved the machine directly to the 10¢-state. One can proceed to build up a picture for the remaining states. The full state-transition diagram for the 25¢-Coke machine is in Fig. 2d. A technical detail is that in a real Coke machine one would like the machine to automatically return from each of the "accept" states to the initial state. This can be done with *e-moves* in our state-diagram. We did not show them because the diagram would have been too cluttered. In essence an *e-move* is that on receiving the "empty" or null token, the machine automatically transfers to another state. Adding e-moves to FA do not change their computational power. Thus by adding an e-move from each of the final states back to the start state, the Coke machine is ready for the next customer.

We note in Fig. 2 that a three token alphabet (N, D, Q) leads to 10 states. Thus, working on the Coke machine problem with an FA, doesn't reduce the complexity of the problem. Instead, the FA provides an orderly mechanism to encode the problem and analyze or study the complexity of the problem. For example, one could use the FA in Fig. 2 to generate the possible sequences of coins leading to a Coke. This of course is a trivial example, but there are many non-trivial examples.

APPLICATION OF STATE-BASED AUTOMATA TO THE GENE-PROTEIN COMPLEX: GENEREAL FEATURES

The gene-protein complex as an LBA. Assume that the current *state* of the system can be identified (1) with the proteins produced by the expressed genes and (2) with the compounds in the cell (substrates, products, etc.) affected by the proteins. Furthermore, assume that *read-write tapes* can be identified with the genome, other internal non-proteinaceous compounds (sugars, etc.) that affect proteins, and the external factors (proteins or non-proteinaceous compounds, temperature, etc.) driving the cell's processes. We assume that the tapes are finite, i.e., we will only consider a finite simulation time. We suggest that the gene-protein complex part of the cell can be treated as an LBA. State-based automata provide a natural description of the gene-protein complex because they can be built around the quantities (protein occurrence) that are directly observable by experiments. Many states of an LBA may contribute to the logic of the biochemistry without including values for all the compounds of interest. These LBA states are precursors to the LBA states that include all the results of the biochemical logic for given time step. We identify all the states that occur at the end of each time period and that contain a complete description of all the constituents as "final states" or "accept states" in the formal definition of an LBA. We refer to all the other states variously as "logical", "calculational", or "intermediate" states. We regard each of our final states as quasi-steady state. That is, each time interval, corresponding to one tape cell being read, is assumed to last long enough for the system to approach the new state. This is very similar to the approach of Thomas and D'Ari (1990).

While some simulation approaches have "state variables", which change continuously in time, we chose to regard each of our states as quasi-steady state. That is, each time interval, corresponding to one tape cell being read, is assumed to last long enough for the system to approach the new state. The assumption of quasi-steady states simplifies the discussion (and the state diagrams) considerably.

APPLICATION OF AN LBA TO THE LAC OPERON: A SPECIFIC EXAMPLE

Sokhansanj et al. fuzzy system model of the lac operon. Sokhansanj et al. (2002) have developed a fuzzy logic model of the lac operon and its regulation. They discuss the problem of gene regulation in some detail and give an exposition of the concepts of fuzzy logic and its application to biological problems in general.

One of the hallmarks of biological problems is uncertainty, which fuzzy logic is particularly well suited to describe. In particular Sokhansanj et al. describe the application of fuzzy logic to the lac operon in which they give a detailed description of rules for determining the outcomes of biochemical interactions. They use the technique of Union Rule Configuration (URC) as introduced by Combs and Andrews (1998) to define their model. Their URC rules are given in their Fig. 9. We shall make use of their discussion as a starting point.

It should be noted that fuzzy systems are well suited to control applications because they produce smooth behavior in the control system or regulator. Thus fuzzy systems have excellent prospects for reproducing any smooth control behavior exhibited by the lac operon. See Kosko (1993) for a very accessible discussion of fuzzy systems.

An LBA model of the lac operon. As an example of applying an automaton to a portion of the gene-protein complex, let us use an LBA as an abstract model for the lac operon. The abstraction of the lac operon used here differs from that of Sokhansanj et al. in two important respects.

First, for all variables of the LBA, let us assume each variable can take on only two possible values; depending on the variable these values are High-Low (glucose, P(lacZY), lacZ, lacY), High-Zero (lactose), Normal-Engineered (P[lacI]), or Normal-Extreme (lacI). We allow two exceptions. See the caption of Fig. 3. The goal of this exercise is to explore the state behavior of the system, i.e., mimicking the major state transitions of the system with a few simple rules. This is to be done by using the minimum of number of states in the state transition table (or graph). We emphasize that using binary tokens is for convenience, i.e., to keep the state transition diagram simple and yet to capture major transitions. The LBA formalism can accommodate tertiary, quaternary, or any number of tokens for a given variable if the data requires it.

Secondly, membership of a variable in either of the two sets is classical, i.e., crisp rather than fuzzy. Thus we anticipate abrupt transitions will occur rather than the smooth transitions produced by a fuzzy system. Our intent is

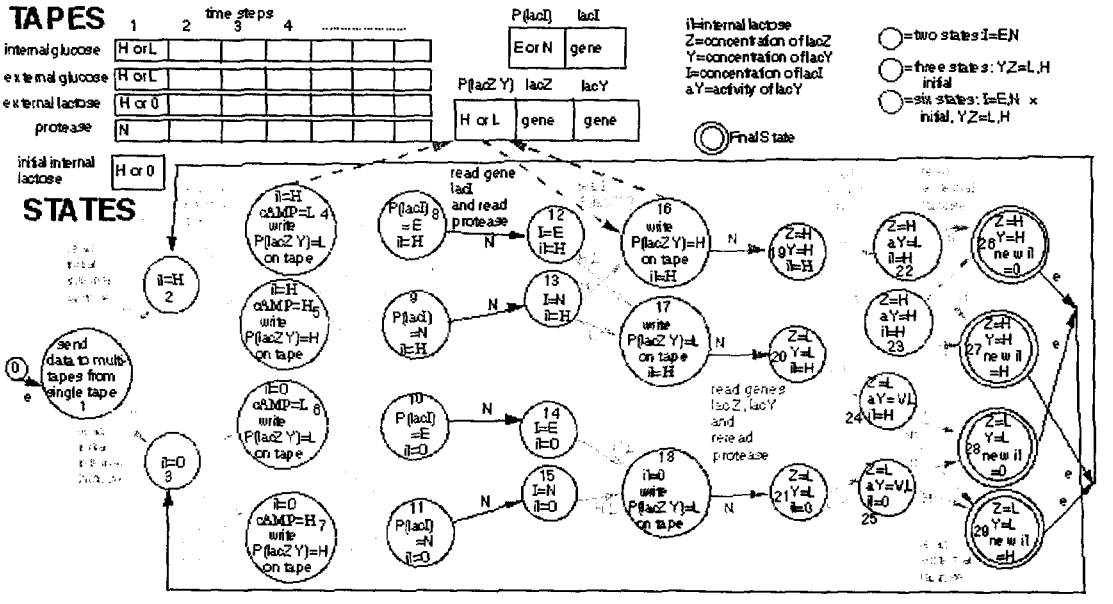


Fig. 3. Schematic diagram of a linear-bounded automaton model of the lac operon. Input driver variables of internal glucose, external glucose, external lactose, and protease are assumed to be placed on two-way read-write tapes with each input value assumed to be for one time step. The status of the promoters P(lacI) and P(lacZY) are also read-write values. States of the LBA are shown as circles. Each state is numbered. States are defined by levels of protein or substrates (lactose). For this model each variable is binary, except for activity of lacY, which can have three values: VERYLOW (V), LOW (L), and HIGH (H), and except for protease concentration, which is always taken to be N (normal). In state 13, the lac repressor is bound to an inducer, allolactose. The repressor is not bound to an inducer in states 14 and 15 and is active.

not to reproduce fine gradations in system control. Our intent is to explore large qualitative changes in the system.

Description of the LBA model of the lac operon. The map and the territory. In describing the operation of the LBA shown in Fig 3, we need to differentiate two different systems: the computations of the LBA and the operation or logic of the underlying real biological system. In describing the transitions of the LBA, we will find that most transitions or states derive from the underlying system.

However a few of the transitions or states come from the nature of the LBA itself.

Filling the tapes in Fig. 3. State 1. The transfer from the start state 0 is an e-move (see p. 5) to state 1. State 1 is a subroutine comprising many sub-transitions and sub-states that fill in the tapes shown in Fig. 3. The formal definition of an LBA posits that there is one tape (not shown) of limited size on which all the input symbols are placed. Without loss of generality, our LBA will read this tape and transfer the results to the set of multiple tapes shown in Fig. 3. To do this we can reserve two special binary symbols for each variable. Thus the machine can differentiate a HIGH for one variable (say internal glucose) from the HIGH for another variable (say external lactose). We shall assume that the LBA is to simulate the behavior of the operon over a finite time period, and we shall assume that the time sequence for the values of any one variable are entered on the input tape in correct order. In state 1, the LBA reads the first symbol on the input tape, recognizes the variable type, and places the symbol in the first blank spot on the tape for that variable. This is repeated in state 1 until the end of the input tape is reached. At that point, the multi-tapes shown in Fig. 3 are filled.

Logic of the lac operon and its abstraction. The lac operon controls internal lactose inside the cell, and, as we shall see, there is a feedback from the current internal lactose level to the operon. Thus the value for internal lactose in the next time step will depend on the new operation of the operon, but the new operation of the operon depends on the current value of internal lactose. So to get the ball rolling as it were, we need to input an initial value of internal lactose. This value can be viewed as what is there at time zero. On reading the initial value of internal lactose, the LBA transits from state 1 to either state 2 or state 3 depending on the value read in.

Lodish et al. (1999) review the lac operon and note that high levels of cAMP only occur when glucose is low. Whenever glucose is high, cAMP levels are low. The cAMP-CAP complex activates the P(lacZY) promoter by binding to the CAP site and stabilizing the binding of RNAP to the promoter DNA. Thus we read in the internal glucose and if it is LOW we set cAMP HIGH and P(lacZY) HIGH; if internal glucose is HIGH we set cAMP LOW and P(lacZY) LOW (states 4 through 7). Let us treat P(lacZY) as part of the tape rather than treating P(lacZY) as a state. Therefore, the LBA writes the effect of internal glucose on P(lacZY) to tape (to be accessed later).

Sokhansanj et al. (2002) point out that the lac repressor promoter, P(lacI), is very weak. This promoter promotes the production of lacI (lac repressor), which binds to an operator of P(lacZY) and reduces production of lacZ (lac enzyme) and lacY (lac permease). We follow the lead of Sokhansanj et al., who included various levels of P(lacI) in their fuzzy system model by assuming its strength was amenable to genetic engineering. We will allow P(lacI) to have two values, either its normal, weak value N or a very high, engineered value E. The LBA reads the P(lacI) site and moves to the appropriate state (8 through 11). In these latter states, the LBA simultaneously reads the gene for lacI and the

protease concentration. LacI reaches a steady state concentration when the production set by the gene is balanced by the destruction controlled by protease (states 12 through 15). In these latter states, P(lacZY) is accessed again to determine the effect of the level of lacI on the system.

Lodish et al. (1999) note that normally if lactose is present (HIGH), a related compound allolactose (Alberts et al. 1983) is an inducer and binds to the lac repressor (state 13) such that the repressor does not bind to the P(lacZY) promoter. Thus, transcription is not blocked if lactose is HIGH. In state 13, the lac repressor is bound to an inducer. State 13 transfers to state 16 if P(lacZY) was previously set HIGH by activation by cAMP. However if P(lacZY) was already set LOW by cAMP being LOW, then P(lacZY) will remain LOW (transition from 13 to 17).

If no lactose is present (ZERO), the lac repressor binds to the promoter. Thus, the repressor in states 14 and 15 is active. In this case, no matter how P(lacZY) was set previously by cAMP, states 14 and 15 transfer to state 18. If there is an extremely high amount (E) of lacI, then no matter what the level of P(lacZY) fixed previously or the level of internal lactose, we assume P(lacZY) is repressed to a low level (transitions 12 to 17 and 14 to 18). Note that states 8, 10, 12, and 14 are hypothetical and are suggested for exploratory purposes. Next the LBA simultaneously rereads protease and reads genes for lacZ and lacY, producing lacZ (lac enzyme) and lacY (lac permease), respectively, in the appropriate amounts. Once again, production balances destruction by protease at steady state (states 19 through 21). Note that lacI imperfectly represses P(lacZY) and thus there is always at least a small amount of lacY and lacZ in the system (Sokhansanj et al. 2002).

Now read in external glucose. External "glucose inhibits the effect of lac permease" (lacY) "at the membrane" (Sokhansanj et al. 2002). This is shown in their Fig. 6. So if lacY is HIGH (state 19) or LOW (states 20 and 21), then we assume a HIGH reading of external glucose produces a lacY activity of LOW (state 22) or VERY LOW (states 24 and 25), respectively. If external glucose is LOW, then the lacY activity is the same as the lacY concentration (states 23, 24, and 25).

So far in my reading of the literature, there is some uncertainty as to what permease activity is so low, which combined with lac enzyme activity, that interior lactose levels are effectively ZERO. The URC rules of Sokhansanj et al. (2002) are consistent with the possibility that all permease levels (V, L, and H) allow enough to enter such that interior lactose is

Table 1. Simulation logic of the lac operon LBA. We assume the three input streams are fixed at the values given for many periods. The initial value for internal lactose is shown for each input stream followed by the results of the LBA simulation for internal lactose for the first three periods. Also shown is whether the operon is ON or OFF and which states the system is in after all inputs are read.

Internal glucose	External glucose	External lactose	Internal lactose value Operon status: State			
			Initial value	Period 1	Period 2	Period 3
0	0	0	0	0 OFF: 28	0 OFF: 28	0 OFF: 28
			1	0 ON: 26	0 OFF: 28	0 OFF: 28
0	0	1	0	1 OFF: 29	1 ON: 27	1 ON: 27
			1	1 ON: 27	1 ON: 27	1 ON: 27
0	1	0	0	0 OFF: 28	0 OFF: 28	0 OFF: 28
			1	0 ON: 26	0 OFF: 28	0 OFF: 28
0	1	1	0	1 OFF: 29	1 ON: 27	1 ON: 27
			1	1 ON: 27	1 ON: 27	1 ON: 27
1	0	0	0	0 OFF: 28	0 OFF: 28	0 OFF: 28
			1	0 OFF: 28	0 OFF: 28	0 OFF: 28
1	0	1	0	1 OFF: 29	1 OFF: 29	1 OFF: 29
			1	1 OFF: 29	1 OFF: 29	1 OFF: 29
1	1	0	0	0 OFF: 28	0 OFF: 28	0 OFF: 28
			1	0 OFF: 28	0 OFF: 28	0 OFF: 28
1	1	1	0	1 OFF: 29	1 OFF: 29	1 OFF: 29
			1	1 OFF: 29	1 OFF: 29	1 OFF: 29

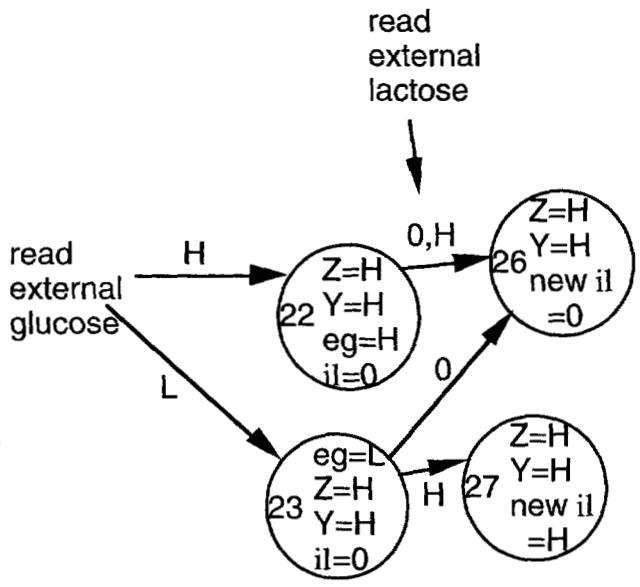


Fig. 4. State transition diagram for the LBA under the assumption of an H read for external lactose would reduce lactose content to ZERO if external glucose severely inhibits lac permease.

Table 2. New simulation logic under change that an H read for external lactose in state 22 leads to a transition to state 26 (Fig. 4) and that all other transitions in Fig. 3 remain unchanged. All other entries in Table 1 remain unchanged.

Internal glucose	External glucose	External lactose	Internal lactose value Operon status: State Initial value	Period 1	Period 2	Period 3
0	1	1	0	1 OFF: 29	0 ON: 26	1 OFF: 29
			1	0 ON: 26	1 OFF: 29	0 ON: 26

nonzero if external lactose is nonzero. This is shown in our Fig. 3. On the other hand if one is to assume that VERYLOW activity of permease is so low that even low levels of lac enzyme can metabolize lactose to effectively ZERO, then the state diagram must be altered somewhat. In this case, the diagram for the transitions from state 19 to states 26 and 27 as drawn is correct. However for this second possibility, states 24 and 25 should each be subdivided into two states (24 [aY=L] and 24a [aY=V], and 25 [aY=L] and 25a [aY=V]). In this case, states 24a and 25a transition to state 28 no matter what external lactose is read in; states 24 and 25 both transfer to states 28 and 29 on 0 reads and H reads, respectively. A third possibility is discussed below in the Analysis section.

In Fig. 3, read in external lactose (ZERO or HIGH). Keep in mind that for this variable HIGH means nonzero. If permease activity allows lactose to enter, then internal lactose will equilibrate to ZERO or HIGH if external lactose is ZERO or HIGH, respectively. Thus states 22, 23, 24, and 25 either go to states 26, 26, 28, and 28 or to states 27, 27, 29, and 29, respectively, depending on whether ZERO or HIGH external lactose is read. These results are an application of the quasi-steady state assumption.

States 26 and 28 feed back to state 3 with an e-move, and states 27 and 29 feed back to state 2.

RESULTS

First consider some preliminary analyses of Fig.3, and then some generalizations suggested by Fig. 3.

Final state. According to the discussion on correspondences in applying an LBA to the gene-protein complex (page 4), the final states or accept states are states 26 through 29. These are the quasi-steady-state states with end products produced by all the biochemical processes for that time step.

Periodicity. It is not easy to examine Fig. 3 visually and decipher all its logical implications. However one can note some general features. First, as long as the input streams continue to be read, the LBA will continue to make transitions to different states. Since there are a finite number of states this must mean that there are cycles in which the machine comes back to a state it had been in before if we give it a long enough input stream. We can ask a series of questions about these cycles. For example: How many cycles are there? If we define a *period* to be the transitions from states 2 or 3 to states 26 through 29, then do all cycles have one period? Are there some cycles in which the LBA has to go through two or more periods before it repeats itself?

Simulation logic of the LBA. To clarify periodicity and final state issues, let us examine the simulation logic of the LBA. Consider input streams in which each input variable is constant over time. There are three input streams to be considered: internal glucose, external glucose, and external lactose. For now we will ignore the ENGINEERED case of P(lacI) and assume the input for P(lacI) is N. In Table 1, we show the simulation logic of the LBA, including whether the operon was ON for the period (state 16 was visited) or OFF (either state 17 or state 18 was visited). We also show which of the final states 26 through 29 were visited. *These latter four states contain a complete description of all the relevant constituents that would be experimentally observed, and they are the result of all inputs.*

Table 1 shows that all inputs produce stable 1-period cycles. In all cases, the stable 1-period cycles are independent of the starting conditions (initial value of internal lactose).

If either external lactose is ZERO or internal glucose is HIGH, then the operon is OFF. If external lactose is ZERO, each 1-period cycle ends in the state 28 final state (LOW lac enzyme, LOW lac permease, ZERO internal lactose).

If initial internal lactose is HIGH, it is possible for the system to go through an intermediate state in which the operon is ON for one period before turning OFF permanently (i.e., input=[0,0,0] or [0,1,0] where [X,Y,Z] designates [internal glucose, external glucose, external lactose]).

If internal glucose is LOW, then external lactose of HIGH always produces an ON operon and HIGH levels of internal lactose.

A curious 2-period cycle. Fig. 3 shows a transition from state 22 to state 27 when HIGH level of external lactose is read in. We assumed in this case that, even though external glucose was HIGH thereby reducing the activity of lacY, enough lactose entered the cell such that the combined value of the entering lactose plus the HIGH amount that was already there was sufficient to insure that some lactose would remain even with a HIGH level of lacZ. While the previous assumption of a transition from state 22 to state 27 is consistent with the fuzzy system model of Sokhansanj et al., one might assume that such a small amount of lactose would enter under these conditions that lacZ could metabolize the entering lactose plus the HIGH amount already there. Then instead of the transition from state 22 to state 27 for an H read, we would have a transition from
state 22 to state 26 with an H read. That is, states 22, 23, 26, and 27 would look like Fig. 4. If one assumes that the transitions from state 24 and from state 25 remain unchanged, then having a transition from state 22 to state 26 under HIGH external lactose produces a change in the [0,1,1] entries as shown in Table 2. All other entries in Table 1 remain unchanged. In Table 2 we see that a 2-period cycle is produced no matter what the initial starting state is for internal lactose. The two different initial values produce the same 2-period cycle, which differ from each other only by a shift of one period. This oscillating behavior raises the question of whether periodic behavior is ever observed in any gene-protein system.

It is not surprising that a change in the model produces a different result. It is reassuring that a change in model structure produced a change in only one sequence out of eight.

The correlation of internal glucose with external glucose. We have been treating external glucose and internal glucose as forcing or driving variables, which are independent of each other. Table 2 forces us to ask if this assumed independence is realistic. If external glucose is highly correlated with internal glucose, then Table 1 is shorter by a factor of 2 and we only need to consider those entries in which external glucose is equal to internal glucose. In fact, an examination of either Fig. 3 or Table 1 reveals that, under the assumptions used to construct Fig. 3, the final state, which is reached, is independent of external glucose. In fact, under the assumptions used to construct Fig. 3, the external-

glucose-read and states 22 through 25 can be deleted without changing the results of the computations.

ANALYSIS AND DISCUSSION

The function of protease. From the point of view of an LBA programmer, protease provides a vehicle of protein destruction to balance protein production. This means that when the system returns to the beginning of the next period of inputs, it eventually "forgets" what states it used to be in as far as proteins are concerned. Between the action of protease and the assumption of time intervals sufficient to approach quasi-steady state, leaving protein in final states is assumed to not affect the next time interval. This suggests the possibility either protease is produced by an unregulated gene or a minimum level of protease is maintained at a non-zero level except for unusual circumstances.

Time steps. We assume our time step to be the overall response time or relaxation time of the system. Implicitly we are assuming an underlying biochemistry that is "taking care of the details" for us. We have structured the LBA for the purpose of describing the occurrence of changes in the internal state of the lac operon. Had our goal been the dynamic approximation of decay of proteins within the context of an LBA, we would have analyzed and compared the relaxation time of protein metabolism to the relaxation time of the other processes. Representing a dynamic approximation to protein decay within an LBA probably has several solutions. One solution might involve the addition of more states in which protein levels were more finely graded. Another solution might be to use the tape to store the current floating point representation. However given the goal here of exposition of simulation of major changes in state by an LBA, we have chosen to fix the protein decay process to one time steps. For steady state applications, this latter approach might be sufficient even for production (non-expository) purposes. This general topic of time scales of processes deserves careful attention in all future applications.

Reaching the accept states. The final states or accept states are designated as states 26 through 29. We reiterate that these final states of the LBA contain specifications of the levels of all the constituents that are observable. The designation of a final state is necessary to meet the formal definition of an LBA. Table 1 tells us about the accept states or final states that is the result of each combination of inputs. If we would like the machine to "end" in either states 27, 28, or 29, i.e., go to a particular stable cycle, then Table 1 tells us which possible combinations can be put together which do that. However, for the LBA to reach state 26 for which an "accept" occurs, one has to stop the machine after one cycle or one has to contrive a complicated sequence. It would appear from Table 1 that in practice state 26 is only entered through unusual circumstances. These results suggest that state 26 might only be entered when the input streams are going through transitions from one set of environmental conditions to another.

Genetic engineering of the strength of promoter P(lacI). Consider the simulation logic of constant inputs if P(lacI) is read in as E. In this case, the operon is always OFF. If the external lactose is HIGH, then a stable 1-period

cycle is established which goes through state 29 with lactose remaining HIGH. If external lactose is ZERO, then a 1-period cycle is established going through state 28. If initial internal lactose is ZERO and if external lactose is HIGH, then a 1-period cycle is established going through state 29. These two results are independent of the initial internal lactose level.

General considerations

Uniqueness. The state transition table (state diagram in Fig. 3) is analogous to a computer program. Hence there is no real uniqueness to the LBA in the sense that two designers could arrive at two LBAs quite different in appearance. However, what is unique is that two identical input streams should produce two identical results in terms of protein levels or substrates. That is, the logic of the two machines should map to each other. How to represent that logic can differ between actual implementations. For example, in the Coke machine automata, we could have had one state for three nickels and another state for one dime plus one nickel. Instead these two states were coalesced into a 15¢ state. However, for finite automata it is known that there is a unique automaton with a minimum number of states for any given language accepted.

Equivalence to Boolean circuit or network? A natural question to consider is whether this approach is a disguised form of a Boolean network (Hatzimanikatis and Lee 1999). If we take binary logic as the hallmark of Boolean networks, then note that our use of mainly binary variables was for convenience and exposition. For example, recall that we had three levels of activity of lac permease. There were four final states. Lodish et al. remark that the low levels of lac enzymes produced by cAMP being low were different from those produced by the repressor being active. We could have differentiated between the two levels and carried that information forward to later parts of the calculation at the expense of using more states. Hence, state-based automata are not restricted to binary logic. In mundane terms, recall that the Coke machine used three tokens.

More generally, are state-based automata equivalent to binary logic circuits no matter what the natural alphabet defined for the automata? A full discussion of this topic is beyond the scope of this report. The power of the LBA to store and process information compared to that of the FA suggests that an FA can be converted to a Boolean network, but in general, an LBA may not be convertible. Boolean networks are fixed like FA are fixed; they are not dynamic like LBA have the capacity to be. This topic should be considered for further study.

Proliferation of states. Logic. Why is there such a profusion of states? There are several factors, which contribute to the profusion. First there is the nature of finite state automata, which the LBA is built on. That is, that the states and the transitions contain the logic of the computation. The state transition table (and graph) is like the "program" of the automaton. If the logic of the program is complicated, then many states can be produced.

Proliferation of states. Combinatorics. If many variables must be "remembered" by the automaton and if differences between states are fixed by the combination of variables, then all the different combination of values can potentially produce an explosion of states.

Proliferation of states. Reading sequentially. The gene-protein complex probably carries out most of its operations in a parallel manner. However an LBA computation of the same processes occurs sequentially by reading in variables from tapes and making decisions. These sequential reads force many states to be generated if there are many different variables or if many accesses of operators on the genome part of the tape must be made. Depending on the logic, the number of states may only increase linearly with the number of "reads".

Information storage. It is much more efficient to store information on the tapes rather than in the states. If there were 10 promoters, each of which were either ON or OFF, this requires 10 locations on a tape. To store this information in states in which each state signifies a different combination of promoters being ON or OFF would require 2^{10} states. This state explosion is one of the advantages of choosing an LBA over an FA. For example, the lac operon could have been simulated with an FA instead of the LBA used here. It would have required more states to do it as an FA.

Larger systems. What generalizations can we draw for larger systems? If, for example, the virulence portion of the genome of some pathogen is 80 genes, then in principle there are $2^{80}=1.2 \times 10^{24}$ different combinations of proteins possible. However in practice, relations between proteins would reduce the number of possible combinations substantially. In other words, we may be able to treat functions within the cell as being conducted by discrete organizational units more or less easily managed by relationships or interfaces. Then it might be possible to have groupings of states in an LBA communicate either through a few states or, if need be, through tape locations. This might be a reasonable way to handle units weakly coupled through only a few connections (e.g. one or two proteins and/or substrates) and which are otherwise independent.

Hierarchical interactions. A possible arrangement that would involve relatively few interactions or connections between proteins and genes is a hierarchical organization. In this scheme, any particular gene would interact with only a few others. The ones, which interact, would form into groups. Thus, any set of genes could be subdivided into groups. In turn, each group would interact with only a few other groups. In this way, the set of groups would be divided into super-groups. At each level of the hierarchy, the current basic entities would form super-entities. Within each super-entity, there would be many interactions (strong). Between super-entities, the interactions would be relatively few (weak). This scheme is very similar to a suggestion by O'Neill et al. (1986) in the ecological literature that organisms (plants and animals) interact hierarchically and communities and ecosystems are organized hierarchically. O'Neill et al coined the term *holon* for the "self-contained" group of interacting entities that become the basic unit entity at the next higher level in the hierarchy.

CONCLUDING REMARKS

We have cast the lac operon onto an LBA framework to demonstrate the feasibility of using LBA's generally to describe gross changes in the state of the protein complex and the associated gene expression.

Construction of an LBA proceeds from a detailed knowledge of the logic of gene regulation and protein production for a particular system. We can arrive at that knowledge by using modern micro-array techniques and finding the set of ODE's that describe the system. One can then extract the logic of the regulation system from the ODE's. Thomas and D'Ari (1990) show how that can be accomplished.

We suggest that a major application of state-based automata will be as submodels in larger models of cell functioning. One reason to construct LBA's from ODE's rather than just retain the ODE's as a submodel of the larger system is to avoid stiff differential equations for the larger system. We suggest that an LBA submodel in a model of cell functioning can determine the mix of enzymes that are present and supply that information to the larger model. That is, the LBA output can be used to define the structure of the equations to be solved in a larger model of cell dynamics.

ACKNOWLEDGEMENT

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

REFERENCES

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson. 1983. Molecular biology of the cell. Garland Publishing, Inc.

Combs, W.R., J.R. Andrews. 1998. Combinatorial rule explosion eliminated by a fuzzy rule configuration. IEEE Trans. Fuzzy Systems 6:1-11.

Cornelis, G.R. 1998. The *Yersinia* deadly kiss. Journal of Bacteriology 180:5495-5504.

Feitelson, D.G.,M. Treinin. 2002. The blueprint for life? Computer 35(7):34-40.

Hatzimanikatis, V., K. H. Lee. 1999. Dynamical analysis of gene networks requires both mRNA and protein expression information. Metabolic Engineering 1: 275 -281.

Hopcroft, J.E., J.D. Ullman. 1979. Introduction to automata theory, languages, and computation. Addison-Wesley.

Kosko, B. 1993. Fuzzy thinking. Hyperion.

Liang, S., S. Fuhrman, R. Somogyi. 1998. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symposium of Biocomputing 3:18-29.

Lodish, H.L., A. Berk, S.L. Zipursky, P. Matsudaira. 1999. Molecular cell biology. W.H. Freeman & Co.

O'Neill, R.V., J.B. Waide, D.L. DeAngelis, T. Allen. 1986. A hierarchical concept of ecosystems. Princeton University Press.

Sokhansanj, B. A., Garnham, J. B., Fitch, J. P. 2002. Interpreting data from microarray experiments to build models of microbial genetic regulation networks. P.27-37. *In* SPIE Photonics West: Biomedical Optics and Applications (BiOS 2002), San Jose, CA, Jan. 19-25, 2002; *Proc. SPIE*, Vol. 4623, Functional Monitoring and Drug-Tissue Interaction.

Somogyi,R., S. Fuhrman, X. Wen. 2001. Genetic network inference in computational models and applications to large-scale gene expression data. p. 119-157. In Computational Modeling of Genetic and Biochemical Networks (J.M. Bower, H. Bolouri, eds.) MIT Press: Cambridge, MA.

Taylor, R.G. 1998. Models of computation and formal languages. Oxford University Press.

Thieffry, D., R. Thomas. 1998. Qualitative analysis of gene networks. Pacific Symposium on Biocomputing 3:77-88.

Thomas, R. and R. D'Ari. 1990. Biological feedback. CRC Presss.

Turteltaub, K., F. Milanovich, P.Fitch. 2002. Pathomics: A national security motivated S&T initiative to redefine biodefense. LLNL White Paper.

UCRL-ID-152287

Preliminary Analysis of Gene Expression Data from Glycolysis in Yersinia Pestis: Application of a Prototype Genetic Algorithm

J. R. Kercher, J. N. Quong, A. A. Quong, C. F. Melius, B. A. Sokhansanj, P. P. Gu, E. Garcia, V. L. Motin

February 26, 2003

U.S. Department of Energy
Lawrence Livermore National Laboratory

DISCLAIMER

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

Preliminary analysis of gene expression data from glycolysis in *Yersinia pestis*: Application of a prototype genetic algorithm

J.R.Kercher*

J.N. Quong†

A.A. Quong¶

C.F. Melius†

B.A. Sokhansanj¶

P.P. Gu§

E. Garcia¥

V.L. Motin¥

*Environmental Sciences Division, L-396
Energy and Environment Directorate

†Chemistry and Chemical Engineering Division
Chemistry and Materials Science Directorate

¶Analytical and Nuclear Chemistry Division
Chemistry and Materials Science Directorate

§Electronics Engineering Technologies Division
Electronics Engineering Department

¥Biodefense Division
Biology and Biotechnology Research Program

Lawrence Livermore National Laboratory
Livermore, California 94550

INTRODUCTION

Regulation in genetic networks can be modeled using ordinary differential equations (ODE's) (Voit 2000, Gibson and Mjolsness 2001). We have developed a prototype genetic algorithm (GA) for automating the determination of the optimal set of ODEs describing a genetic network. Genetic algorithms were originally developed by Holland (1962, 1973, 1975), have since been extensively studied and reviewed (Goldberg 1989, 2002; Haupt and Haupt 1998, Mitchell 1996, Reeves 1993), and have been used in many applications including bioinformatics (Fogel and Corne 2003), oil field management (Johnson and Rogers 2001), optimization of groundwater decontamination (Rogers et al. 1995), and analysis of metabolic pathways (Koza et al. 2000).

We have conducted an exercise to analyze data from gene expression of glycolysis genes of the *Yersinia pestis* bacillus without the use of *a priori* knowledge of identities of genes or proteins. In this exercise, we have found candidate systems of ordinary differential equations whose solutions are in good agreement with the data. The goal of this exercise was to determine what results were possible using no information other than that contained in time-series data of gene expression itself.

EXPERIMENTAL DESIGN

Yersinia pestis was grown at 26 degrees Celsius for 48 hours. At that time, the temperature of some colonies was elevated to 37 degrees Celsius. Gene expression data was collected at one hour, four hours, and ten hours post elevation of temperature. The results were expressed as the ratio of gene expression (mRNA concentrations) for the 37-degree samples to those of the 26-degree samples. The raw data is shown in Table 1. The negative signs indicate that the ratios are actually of the 26-degree samples to the 37-degree samples. We modify this data to remove the negative signs and put all data entries as a ratio of the expression at 37 degrees to that at 26 degrees. This is shown in Table 2.

TWO ALTERNATIVE METHODS TO DESCRIBE THE NETWORK

Representing the Network with Discrete States

Some authors have argued that for certain purposes it is appropriate to represent the behavior of genetic networks by discrete-state schemes rather than by continuous variables, such as solutions to ordinary differential equations (Thomas and D'Ari 1990, Gibson and Mjolsness 2001, Somogyi et al. 2001, Kercher 2003).

The simplest discrete-state representation. If we are interested in phenomena that have much longer characteristic times than that required for re-equilibration of the gene expression to the new temperature regime, then the we can assert a simple description of the temperature response using the data in Table 2. Gene expression changes state over a 10-hour period such that, by the end of the period, genes G, I, L, and M are up-regulated and gene Q is down-regulated. The rest of the genes have returned to their original expression by the end of the ten-hour period. Here we are using the criteria that a net change of less than plus or minus 20 percent is no change.

This result is purely descriptive. During the ten-hour period, gene expression for almost all of the genes has undergone significant transitory excursions away from their steady state values before coming back under control. Ideally one would like an

Table 1. Raw microarray data for *Y. pestis*. Relative increase in gene expression for *Y. pestis* when grown at 37 vs 26 degrees at three time points. The minus signs indicate the inverse ratio.

Gene Code	Hr 1	Hr 4	Hr 10
A	-1.15702	-1.26041	-1.17197
B	-1.86144	1.094229	-1.08293
C	-1.60908	-1.11398	1.088521
D	-1.27982	-1.48129	-1.17857
E	-2.19081	-1.21216	-1.13406
F	-1.53606	1.062806	1.123661
G	-3.26023	-1.32626	1.313621
H	-1.12926	-1.02536	1.131597
I	-1.41479	1.20661	1.416591
J	-1.82555	-1.0607	-1.08627
K	-2.02379	1.050368	1.0968
L	1.102746	-1.04538	1.534276
M	-1.25508	1.059375	1.56482
N	1.463858	1.139525	-1.1184
O	-3.02518	-1.34833	-1.14903
P	-1.19202	1.052258	1.03045
Q	3.99545	1.083866	-3.49379
R	-2.51384	-1.44457	1.001729
S	1.30433	1.147458	1.093641
T	-1.60908	-1.11398	1.088521
U	-1.17419	-1.3077	-1.11107
V	2.292165	-1.02039	1.078728

understanding of the signals being processed by the interconnected network of genes and proteins that can produce such behavior.

A more detailed state-based description. Thomas and D'Ari (1990) propose a scheme of finite states in which such transitions can be described. Under their scheme, one would include the intermediate state in which B, C, D, E, F, G, I, J, K, M, O, R, and T are down-regulated at one hour, the genes N, S, Q, and V are up-regulated at one hour, and the others remain the same at one hour. In the scheme of Thomas and D'Ari such a state would be both describable and metatstable under time-development and would transform into the final state described above.

The scheme of Thomas and D'Ari is a simplification or abstraction of ordinary differential equations. They do not give a systematic method of arriving at the differential equations or the states given an arbitrary time-series data set, especially for a large number of genes expressed. Thus, finding the set of ordinary differential equations that govern the network of 22 genes, is the first step in abstracting them to the finite state description of Thomas and D'Ari.

Representing the Network by Ordinary Differential Equations

In addition to the fact that the discrete-state system of Thomas and D'Ari is founded on first determining the parameters for a continuous variable representation, there are

Table 2. Data from Table 1 revised so that all data are the ratio of expression at 37 degrees to that at 26 degrees.

Gene Code	Hr 1	Hr 4	Hr 10
A	0.864287	0.793394	0.853266
B	0.537218	1.094229	0.923422
C	0.621472	0.897683	1.088521
D	0.781358	0.675088	0.848486
E	0.456453	0.824973	0.881784
F	0.651018	1.062806	1.123661
G	0.306727	0.754001	1.313621
H	0.885537	0.975265	1.131597
I	0.70682	1.20661	1.416591
J	0.547779	0.942775	0.920581
K	0.494124	1.050368	1.0968
L	1.102746	0.956586	1.534276
M	0.796759	1.059375	1.56482
N	1.463858	1.139525	0.894131
O	0.330559	0.741656	0.870299
P	0.838912	1.052258	1.03045
Q	3.99545	1.083866	0.286222
R	0.397798	0.692247	1.001729
S	1.30433	1.147458	1.093641
T	0.621472	0.897683	1.088521
U	0.85165	0.7647	0.900032
V	2.292165	0.980013	1.078728

several other reasons why a continuous variable approach, i.e., ODE's, is useful to analyze time-series data of the type in this exercise. First, if the characteristics times of the problem are large or comparable to the time-scales in the time-series data, then any changes in the data should appear smooth rather than abrupt. In such cases, continuous variables and their accompanying ordinary differential equations are a natural method to describe such data. In fact, the data in Table 2 is reasonably smooth rather than abrupt. Second, the time-series data gathered by micro-array experiments contain information about the relative sequence of changes in expression. For example, if gene X drops in expression first followed by an increase in expression of gene Y at a later time, that sequence will be shown. Thus any causal-based system, either discrete or continuous, must produce results consistent with the causality implied by the data. The terms in a set of differential equations explicitly contain the causality of the network and the solutions of the ODE's should predict changes in their proper sequence. Furthermore, ODE's, which correctly describe the workings of a genetic network, can be used as the basis for other state-based systems such as the automata approach proposed by Kercher (2003). By finding differential equations that fit the data, we will also be finding the network connections.

Here we report on the exercise of finding sets of differential equations which satisfy known properties of genetic networks and which fit the data. A natural set of questions regarding this exercise are (1) Whether the solutions for the network are unique or are there several representations which can describe the data equally well given the uncertainty in the data. (2) Are the solutions found for the networks consistent with existing knowledge about the glycolysis network. (3) Are good fits artifacts of the ODE's, i.e., is the problem really

under-determined. The answers to these questions go to the issue of whether this type of data is sufficient to determine networks or whether additional experiments are required. Sufficiency of the data will be discussed briefly in the Discussion Section below.

APPROACH TO SOLVING THE INVERSE PROBLEM

We are given a set of data and wish to find the ODE's for the genetic network which best fit the data. We shall refer to this as the inverse problem. Our approach to this problem consists in the following steps. First determine how to represent the structure of the system of equations. To do this we will first define the problem biologically, then mathematically, and then computationally. We will then discuss the computational implementation and the results of the calculation.

The Biological Problem and the Statistics of the Data

In regulated genes, RNA polymerase attaches to the DNA in regions known as promoters and proceed to "process" the group of genes downstream from that promoter. We assume that all the genes in this group of genes belonging to this one promoter are expressed at the same rate. The binding of the RNA polymerase may be facilitated (activated) by other proteins (activators) or may be blocked or in some way interfered with (repressed) by other proteins (repressors). Our first task is to get an estimate of how many promoters there are and which genes might belong to them.

Because we assume that any two genes belonging to the same promoter should be expressed at the same rate, we reason that the expression of any two genes belonging to the same promoter should be highly correlated with each other. Thus the first step we took in analysis was to find the time-series correlation coefficient of all the genes. Those genes, which we found to be highly correlated, were then plotted and visually examined. Visual examination of the resulting plots (Fig. 1) of all the data in Table 2, suggested that the following groups were good candidates for belonging to single promoters: ADU, BE, CT, FK, G, HP, IM, JO, L, NS, Q, R, and V. This suggests that 13 promoters might be adequate for representing the system. Note, in the method described below, we did NOT force any specific gene to be assigned to any specific promoter or force to be included in any specific group with other genes. This part of the exercise only determined that 13 promoters might be adequate. We allowed room for expansion to 16 promoters if necessary.

The Mathematical Problem

Consider a particular protein (with concentration P) produced by (translated from) a particular messenger RNA molecule (with concentration M) generated at a particular gene (available DNA fragment with concentration D_f). Assume there is one activator involved (concentration A). Assuming first order kinetics, we describe the change in concentration of M to be given by

$$\frac{dM}{dt} = c' A \cdot D_f - k_d M \qquad (1)$$

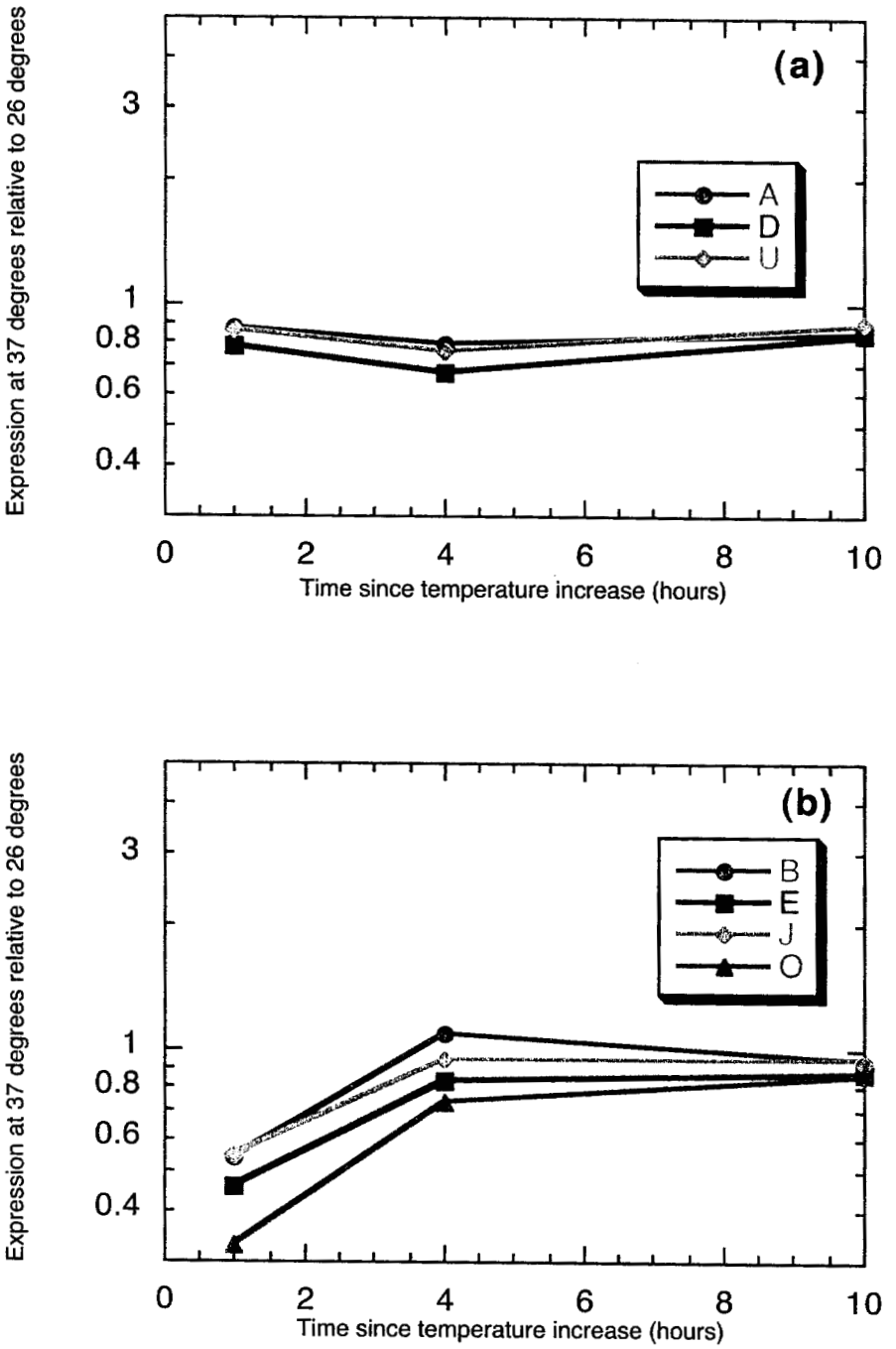
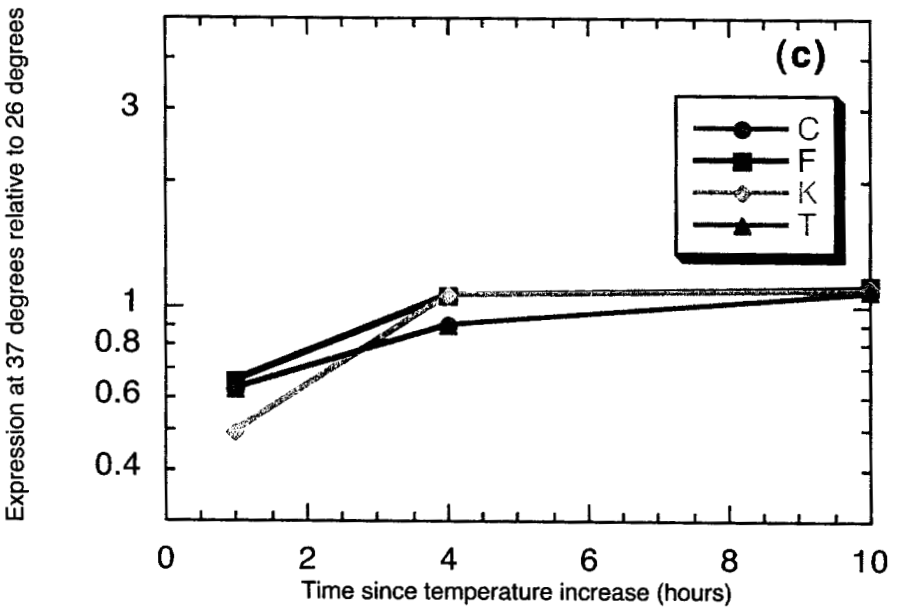


Fig. 1. Microarray exression data of *Yersinia pestis* glycolysis genes. Data shown is ratio of expression at 37 degrees to that at 26 degrees. Genes are labeled alphabetically. Data is sorted into groups to simplify visual presentation.



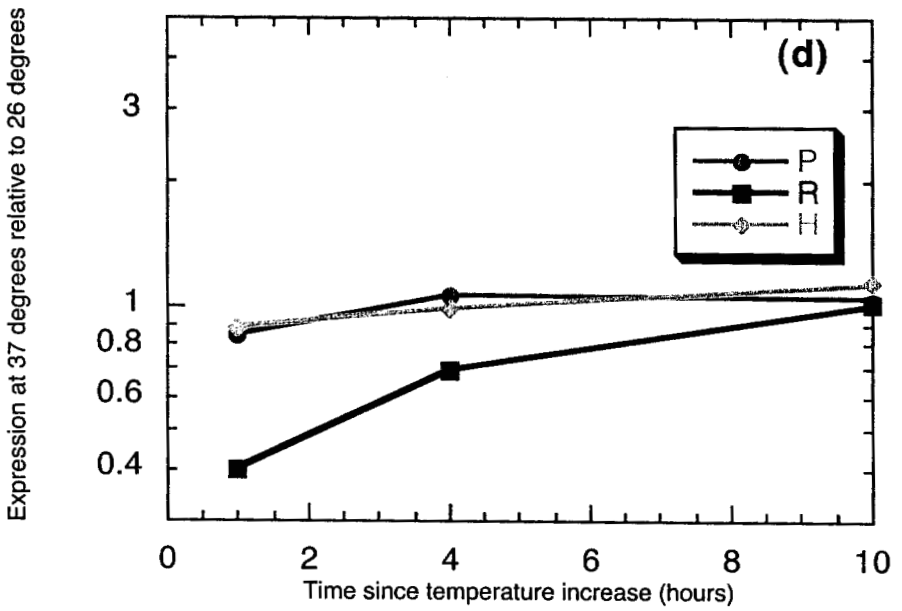
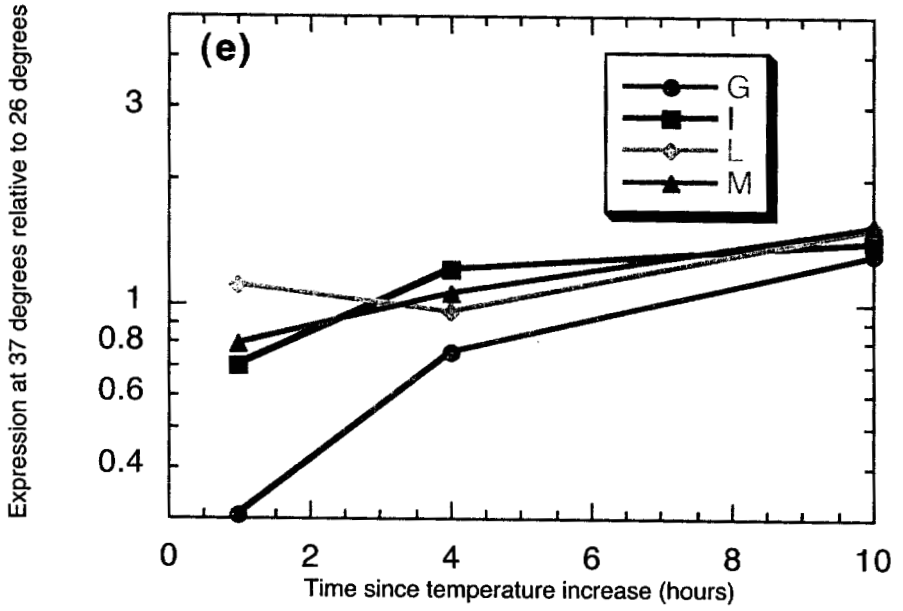


Fig. 1. (continued). Microarray exression data of *Yersinia pestis* glycolysis genes. Data shown is ratio of expression at 37 degrees to that at 26 degrees. Genes are labeled alphabetically. Data is sorted into groups to simplify visual presentation.



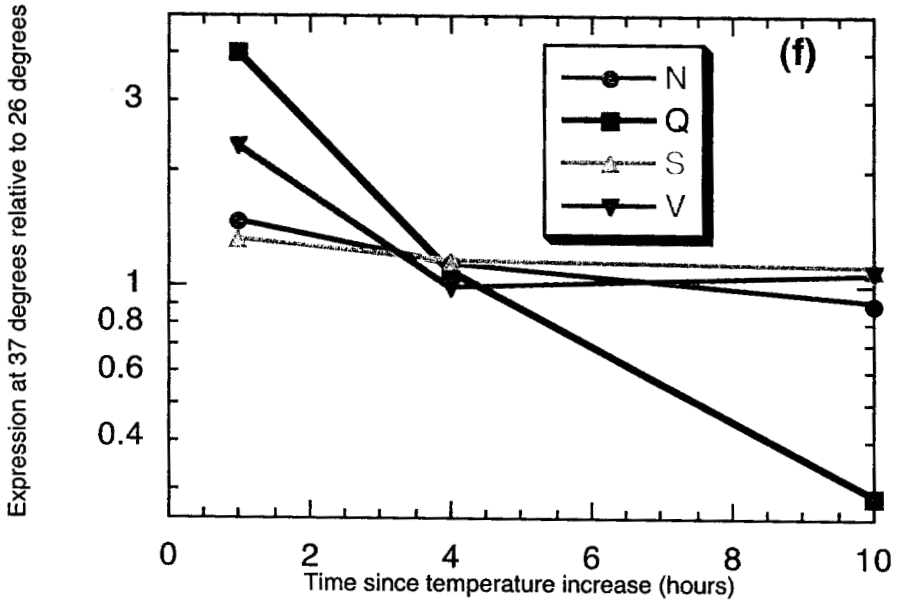


Fig. 1. (continued). Microarray exression data of *Yersinia pestis* glycolysis genes. Data shown is ratio of expression at 37 degrees to that at 26 degrees. Genes are labeled alphabetically. Data is sorted into groups to simplify visual presentation.

where c' is a constant of production and k_d is a decay constant. The subscript f on D indicates free DNA available for binding. The total DNA of this type in the system is given by

$$D_{TOT} = D_f + D_{Bound} = D_f + k_b D_f A \tag{2}$$

where we assume that the bound DNA is proportional to the free DNA and activator concentration. Using eq. 2, eq. 1 can be converted to the usual Michalis-Menten formulation

$$\frac{dM}{dt} = c\frac{A}{1 + c_1 A} - k_d M \tag{3}$$

The equation for the change in protein concentration is given by

$$\frac{dP}{dt} = c_e P_A M - k_p P \tag{4}$$

where c_e is a constant of production, P_A is the concentration of the available amino acids, and k_p is a decay constant. To keep the algebra as simple as possible, we will assume that the amount of available amino acids is held constant independent of the activity of the glycolysis system and so P_A can be absorbed into the constant c_e. We will also assume that the mRNA kinetics is fast compared to the protein kinetics (dM/dt=0) and we will replace M by the Michalis-Menten expression in eq. 3 and add in a small base amount of expression of mRNA, independent of the activator concentration, and hence a small base amount b of protein production. Thus

$$\frac{dP}{dt} = b + c_A\frac{A}{1 + c_1 A} - k_p P \tag{5}$$

where the first two terms are also proportional to mRNA expression. For repressor proteins, (concentration R), similar arguments produce the equation

$$\frac{dP}{dt} = b + c_R\frac{1}{1 + c_2 R} - k_p P \tag{6}$$

If both an activator and a repressor operator is present on a promoter with corresponding activator and repressor proteins, then we approximate the relevant equation by

$$\frac{dP}{dt} = b + c_{AR}\frac{A}{1 + c_1 A}\frac{1}{1 + c_2 R} - k_p P \tag{7}$$

If there are two repressors operators at a promoter, interacting with two distinct repressors of concentration $R1$ and $R2$, then we assume a change in concentration of the product to be given by

$$\frac{dP}{dt} = b + c_{RR}\frac{1}{(1 + c_1 R_1)(1 + c_2 R_2)} - k_p P \tag{8}$$

Note that we are ignoring any possible post-transcription factors, which might lead to non-proportionality between expressed mRNA and protein production.

A Test of the Equation

In a preliminary trial, we considered two coupled equations each of the form of eq.7. We assumed an external protein had a temperature sensitive activity and was the activator of one of the proteins. We found that results of the form of one of individual tracks of expression data could be readily generated. It was readily apparent that to fit all the data simultaneously and account for all of the data and using only internally consistent equations would be exceedingly difficult.

THE PROTOTYPE GENETIC ALGORITHM

In order to solve the inverse problem, we have elected to use a genetic algorithm. In our judgment the key to the genetic algorithm is to find a proper representation of the network structure (colloquially referred to as the "wiring diagram") that is consistent with the genetic algorithm approach.

Summary of the Genetic Algorithm Procedure

We initialize the algorithm by generating an initial population of possible solutions to the inverse problem. That is, each member of the initial population represents a possible "wiring diagram" generated at random and a set of parameters for that wiring diagram, also generated at random. Then the differential equations for each member of the population are solved for 26 degrees Celsius for 48 hours. At that time the temperature is changed to 37 degrees and the solutions are continued forward for one, four, and 10 additional hours. A cost function is calculated by taking the square of the difference of the log of the data less the log of the calculated (model results) ratios of the expressed genes at 37 degrees to the expressed genes at 26 degrees. These residuals are summed over all 22 genes and all three time-points (one hour, four hours, and ten hours). We sort the initial population by costs and discard that half of the population whose members have the highest costs.

Then the genetic algorithm begins a loop where each pass through the loop is one generation. In each generation, the following occurs. First the algorithm selects which members of the population are going to be parents for the next generation. This selection is based on the cost function or fit to the data. Those with the best fit have proportionally greater probability of being parents. Parents are selected in pairs: a "mom" and a "dad". Each pair of parents produce a pair of offspring. The portion of the population (one-half in our algorithm) from which parents are chosen and which are the top performers, remain into the next generation. Offspring replace the members in the lower half of the population; the bottom half performers.

We describe each member by a "chromosome", which is a linear array of parameters that define the network and connection strengths. During "reproduction" a random point is chosen on the chromosome (crossover point) and all parameters to the left of that point for one offspring come from one parent and all parameters to the right of that point come from the other parent. This is reversed for the other offspring.

After the crossover operation is complete, a final operation of mutation is performed. For a fixed mutation rate R, the number of mutations N in the population is given by $N=R$ $\times$(population size)$\times$(number of parameters). The locations of these N mutations are

selected at random in the population's chromosomes. The mutated values are selected at random from the range of allowable values for that parameter.

The above description is a general description of the genetic algorithm, which applies to most cases of its use. Unique aspects of the genetic algorithm used in our application will be discussed next.

Unique Features Of The Prototype Genetic Algorithm

Operators at each promoter. At each promoter we assume there are two operators. We will allow either one activator operator and one repressor operator or two repressor operators.

The real parameter chromosomes. Eq. 7 (or 8) has five parameters. The parameters b, c_m, $c1$, and $c2$ are associated with the promoter and the parameter k_p is associated with the protein produced. The parameter c_m is the generic parameter for protein interaction strength, which denotes c_{AR}, c_{RR}, c_A, and c_R above. Without loss of generality we may set the parameter b equal to 1 because we will be dividing each solution at 37 degrees by the solution at 26 degrees. The solutions will have terms proportional to the terms in eq. 7 with b and c_m and hence we can divide both numerator and denominator of the ratio by b and redefine a new parameter $c_p=c_m/b$. Thus there are three parameters (c_p, $c1$, and $c2$) associated with each promoter and one parameter k_p associated with each protein. We define an array $\mathbf{k2}$ of the k_p with length 22, which is the number of expressed mRNA and is assumed to also equal the number of proteins produced (*maxgene*). That is $\mathbf{k2}=(k_p1,k_p2,....,k_p,maxgene)$ where *maxgene*=22. Likewise there are arrays (chromosomes) for the promoter parameters where the number of promoters is denoted by*nopromoter* and is equal to 13. These additional chromosomes are $\mathbf{k1}=(c1,1, c1,2,....,$ $c1,13)$, $\mathbf{k1}'=(c2,1, c2,2, ..., c2,13)$, and $\mathbf{x2max}=(c_p,1, c_p,2, ..., c_p,13)$.

The range of the real parameters. Genetic algorithms were originally developed for application to problems with binary values of the parameters. They were subsequently extended to problems with integer values and then to problems in which the parameters take on real values. Other search or optimization algorithms can find real parameters on an unrestricted domain. The simplest genetic algorithms assign the values of the parameters from a restricted domain with fixed ranges. For this exercise we have taken this approach. In the results shown here, in the initialization, crossover, and mutation operations, we have limited the values of $\mathbf{k2}$ to the range 0.05 to 5.0 and the values of $\mathbf{k1}$ and $\mathbf{k1}'$ to the range 0.01 to 10.0. We used 0.01 for the lower value of the range of $\mathbf{x2max}$. We performed one set of trials using 10.0 for the upper bound of the c_p in $\mathbf{x2max}$, and a second set of trials using 100.0. The range of $\mathbf{k2}$ was decided based on the time behavior in Fig. 1. For the other real parameters, we assumed one to two orders of magnitude around the value of b should be a sufficient range. In future work on GA's applied to gene networks, adaptive strategies for real parameters in which mutations etc are allowed to occur outside of the initialization range might be considered.

The "wiring diagram" chromosomes are the most difficult part of the data structures. We will turn to them next.

The promoter-membership chromosome. We assume the gene associated with each protein must "belong" to (or be downstream and under at least partial control of) a

promoter. We assume this because the data suggests that each of the 22 genes is under some kind of regulation. So we define a chromosome array of *maxgene* entries called **prombelong**. First, we conceptually assume the promoters are numbered from 1 to *nopromoter*. Then at initialization, each entry in **prombelong** is assigned at random a value between 1 and *nopromoter*. This is the number of the promoter to which the gene/protein "belongs".

The network-connection chromosome. Recall that we assume that there are two operators at each promoter. We need a chromosome, which specifies which protein activates or represses which operator. To accomodate this requirement, we define an array **listprotatop**, which contains 2**nopromoter* entries. We adopt the convention that first pair of entries are entries for the activator and repressor operators of the first promoter, the second pair of entries in **listprotatop** is the entries for the activator and repressor operators of the second promoter, etc. The odd entries are for the activators and the even entries are for the repressors. The value of the entry in **listprotatop** is the number of the protein, which interacts with that operator. These entries are initialized for the first generation. Constraints on how these entries are chosen are discussed below.

Renaming the promoters. In initial trials of the genetic algorithm, we discovered that the algorithm had difficulty with the monstrous combinatorics of the problem as stated thus far. We need to avoid unnecessary degrees of freedom. The random naming of the promoters was a source of unnecessary and potentially harmful extraneous combinatorics. To keep an example simple, suppose that there were seven genes/proteins, which could belong to four promoters. Thus a possible **prombelong** would be (3,4,4,1,3,4,3). Another possible **prombelong** would be (4,3,3,2,4,3,4). Note that these two **prombelong**'s are actually equivalent topologically. Thus we might have two very good solutions, defined by these two chromosomes and their associated chromosomes **k1**, **listprotatop**, etc. These two solutions might actually lie very close to each other and both might be suitable candidates for parenthood for the next generation. However a crossover in **prombelong** would destroy the topology in the offspring and might very well produce two very bad solutions. Hence we need a promoter-naming scheme in which two topologically equivalent **prombelong**'s would have the same entries. The solution we use is that at the end of each generation, just before the cost function is evaluated for the new generation, the promoters in each **prombelong** are renamed with the following convention. The groups of gene/proteins belonging to each promoter are ordered according the value of the lowest numbered member. The number of the promoter that each group of gene/proteins belong to, is fixed as the ordinated number of the group. In our example of seven genes and four promoters, in the first instance the genes/proteins 1, 5, and 7 belong to promoter 3; 2, 3, and 6 belong to 4; and 4 belongs to 1. In the renamed system 1, 5, and 7 would belong to 1; 2, 3, and 6 would belong to 2; and 4 would belong to 3. In the second instance, genes/proteins 1, 5, and 7 belong to 4; 2, 3, and 6 belong to 3; and 4 belongs to 2. In the renamed system 1, 5, and 7 would belong to 1; 2, 3, and 6 would belong to 2; and 4 would belong to 3. Hence in both instances the renamed **prombelong** would be (1,2,2,3,1,2,1). Using these renamed promoters, the values of **listprotatop** and the real chromosomes **k1**, etc. would also be permuted to preserve topology of the network.

Temperature dependence of glycolysis expression. For our initial exercise reported here, we assume that an efficient way for the cell to manage the response to changes in the current value of the temperature is for the activity of a special protein to be extremely sensitive to temperature and that this one protein act as a messenger to all the systems of the cell. Thus we assume for this initial exercise that the activity of an external protein (not one of the 22 in Table 1 and 2) is very sensitive to temperature and activates or represses one or more of the glycolysis proteins, which in turn affect the rest of the glycolysis system. This

protein (23) is treated as a forcing protein in the set of coupled differential equations. The value of the activity is 1 at 26 degrees before the temperature change. The value of the activity is $c_{p,14}$ after the temperature change. In other words, we store the activity of the external protein at 37 degrees in the chromosome **x2max** in the (*nopromoter*+1) position.

We make two remarks about the temperature assumption. First, we are ignoring any sensitivity of the couplings themselves to temperature. This assumption could be a source of error in the computations described below and might make an exact fit of the data impossible. A future exercise would be to relax this assumption and see if the fit improves. One consideration in deciding to ignore temperature dependence of the couplings is that allowing dependence could potentially double the number of parameters and might prove an insurmountable obstacle for the genetic algorithm. At the beginning of the exercise we did not know whether this was the case or not, and so we elected the conservative approach.

The second remark is that it should be a straightforward modification to substitute the assumption that one of the glycolysis proteins (one of the 22) is the one that signals the temperature change rather than an external protein. One would need an "it" parameter to put into a chromosome to determine which protein is "it". Then the reaction strength wherever "it" coupled to an operator could be modified with temperature change.

Two repressors vs an activator and a repressor. A priori we anticipate (hypothesize) that two operators at each promoter are all that will be necessary to describe the data. However, inspection of the Fig. 2, especially Fig. 2f, suggests that there may be some cases in which two repressors at one promoter might perform better and other cases in which an activator operator and a repressor operator at one promoter might be a better description. So we allow the genetic algorithm to have this flexibility. We implement this flexibility by always fixing the even operators (2, 4, ...26) in **listprotatop** as repressor operators. However, the odd operators (1, 3, 5, ..., 25) can be either activator operators or repressor operators. The signal, which carries this distinction is the sign of $c_{1,i}$ in the $\mathbf{k}_1$ chromosome. If $c_{1,i}$ is positive then the protein is taken as an activator, and we use eq. 7 for promoter i. If $c_{1,i}$ is negative, then the protein is regarded as a repressor and we use eq. 8 for promoter i.

If crossover occurs at a position i in $\mathbf{k}_1$ and both parents have positive $c_{1,i}$, then both offspring will have positive $c_{1,i}$. If both parents have negative values of $c_{1,i}$ (repressors), then both offspring will have negative values of $c_{1,i}$. If one parent has a positive value of $c_{1,i}$ and one a negative value, then one offspring will have a positive value and one a negative value.

The crossover rule for real parameters. As stated above, the parameters preceding the crossover site come from one parent; the parameters following the site come from the other parent. For logical parameters (TRUE or FALSE values) and for integers, crossover occurs BETWEEN parameters on the chromosome. However, for real parameters crossover occurs AT a particular parameter. This difference is because real parameters usually adjust the form of existing functions and a blend of values between that of the two parents to the offspring provides the algorithm with a "fine tuning" knob. Integers usually determine the combinatorics and set the form of the functions to be tested.

Assume crossover occurs at real parameter a, (a_{DAD} and a_{MOM}). In general, the parameter a for the two new offspring (a_1 and a_2) is given by

$$a_1 = \beta\, a_{MOM} + (1-\beta) a_{DAD}$$
$$a_2 = (1-\beta) a_{MOM} + \beta\, a_{DAD} \qquad (9)$$

This is the case for chromosomes **x2max** and **k1**'. However, for the chromosome **k1** with entries $c_{1,i}$, we first calculate the absolute values of the offspring parameters by using absolute values of a_{DAD} and a_{MOM} in eq. 9. Next, to determine which offspring parameter gets the minus sign and which offspring gets the positive sign, we calculate the values in eq. 9 using the algebraic values of a_{DAD} and a_{MOM}, not the absolute values. Call these values a_{sign1} and a_{sign2}. Whichever offspring has the lower value of a_{sign} receives the minus sign; whichever offspring has the higher value of a_{sign} gets the positive value.

Restrictions on the interactions in protein regulation of gene expression (the wiring diagram). Protein regulation of genetic expression has many different modes of operation. For example, some genes are self-activating. Some proteins activate many different genes. Other proteins repress many different genes. Some proteins can both activate and repress. Given this range of mode of action and not knowing what possible restrictions might govern genetic regulation in the glycolysis network, we elected to find solutions with several different restrictions on the mode of action of the proteins. There are five classes of restrictions, which are differentiated within the computer code by the integer variable `irestrict`. Table 3 contains the description of the restrictions on protein regulation of genetic expression determined by the value of `irestrict`. For this exercise, we have only enforced the restrictions on the results of the crossover operation and mutation of offspring. We have relaxed these constraints on the results of the mutation operation on parents. The effects of enforcing the restriction on the results of the mutation operation on the parents can be investigated more thoroughly at a later time if warranted.

Implementation of `irestrict`. At the end of each generation, we employ "fixup" routines to modify the offspring to bring them into line with the restrictions in Table 3. For `irestrict=1`, we first fixup any instance in which a protein activates and represses the same promoter. Next we fixup any instance in which either an activator activates more than one promoter or an activator also represses. These instances are "fixed" by changing entries in **listprotatop**. Finally we fix any case in which a protein activates the promoter its gene belongs to. This last is implemented by changing the membership entry of the promoter in **prombelong**. For `irestrict=2`, as before, we change entries in **listprotatop** so that no protein activates and represses the same promoter. We then adjust **prombelong**, if necessary, so that no protein activates the promoter its gene belongs to. For `irestrict=3` and 4, we introduce a new chromosome, denoted by **prottypeact**. The chromosome **prottypeact** is an array of TRUE/FALSE logical values each of which designates whether the protein is an activator or a repressor. An entry of TRUE in position *i* indicates that protein *i* is an activator; a FALSE entry indicates it is a repressor. There are *maxgene*+1 entries in **prottypeact**. The additional entry designates whether the external protein is a repressor or an activator. After each crossover/mutation operation, the new **prottypeact** takes precedence over **listprotatop**. We then adjust **listprotatop** accordingly to bring it into conformity to the restrictions with Table 3.

Model equations. Consider protein *j* produced by gene *j* in the *imember* of the population of chromosomes. It belongs to promoter *i*=**prombelong**(*j*,*imember*). The proteins controlling promoter *i* are given by **listprotatop**. Let

$$k = \texttt{listprotatop}\left(2(i-1)+1, imember\right)$$
$$l = \texttt{listprotatop}\left(2i, imember\right) \tag{10}$$

Then the equation for the rate of change in the concentration of j is given by

$$\frac{dx_j}{dt} = \begin{cases} 1 + \dfrac{c_{p,i}x_k}{\left(1+c_{1i}x_k\right)\left(1+c_{2i}x_l\right)} - k_{p,j}x_j & c_{1i} > 0, k > 0, l > 0 \\ 1 + \dfrac{c_{p,i}}{\left(1+|c_{1,i}|x_k\right)\left(1+c_{2,i}x_l\right)} - k_{p,j}x_j & c_{1,i} < 0, k > 0, l > 0 \\ 1 + \dfrac{c_{p,i}x_k}{\left(1+c_{1,i}x_k\right)} - k_{p,j}x_j & c_{1,i} > 0, k > 0, l = 0 \\ 1 + \dfrac{c_{p,i}}{\left(1+|c_{1,i}|x_k\right)} - k_{p,j}x_j & c_{1,i} < 0, k > 0, l = 0 \\ 1 + \dfrac{c_{p,i}}{\left(1+c_{2,i}x_l\right)} - k_{p,j}x_j & k = 0, l > 0 \\ 1 - k_{p,j}x_j & k = 0, l = 0 \end{cases} \tag{11}$$

Single Chromosome vs Multiple Chromosomes

Genetic algorithms face a daunting task of exploring extraordinarily complicated multi-dimensional "surfaces" to find the global optimum points on these surfaces. Some experimentation and theoretical investigations have gone into studying how well GA's work under various types of approaches. Work to date holds that GA's accumulate good substrings of the chromosome, which are recombined in various combinations by the crossover procedure. If each chromosome is treated independently such that a cross over occurs within each chromosome of a multiple set, then "good" combinations of GA "genes" (model parameters) can get separated from each other very readily. If one the other hand, all of the GA "genes" are in one chromosome, then only one crossover will occur for each pairing and "good" substrings can stay together easier. According to this argument, the best performance is expected if all the multiple chromosomes are concatenated into one long chromosome.

We have elected to try both approaches. So in the Results section to follow, we will report on solutions found using one long chromosome consisting of all the real, integer, and (where appropriate) logical chromosomes. In this approach there is only one crossover point for each pairing of parents. We will also report on the results from the approach of treating each chromosome as independent from the others. In this approach, there is one independent crossover point within each of the six (or seven for `irestrict`=3 or 4) chromosomes.

Table 3. Restrictions on offspring on protein regulation of genetic expression determined by the value of `irestrict`. For this preliminary exercise no restrictions were applied to mutations of parents.

`irestrict`	Restrictions
1	1. No protein activates two or more genes. 2. No protein both activates and represses. 3. No protein activates its own promoter.
2	1. No protein activates its own promoter. 2. No protein activates and represses the same promoter.
3	1. Each protein is either an activator or a repressor. 2. No protein activates its own promoter.
4	1. Each protein is either an activator or a repressor.
5	1. No restrictions.

RESULTS

Cost function. Let $x_{ij}(T)$ denote the observed expression of gene i at time j at temperature T and $m_{ij}(T)$ denote the result of the model for the same variable. The range of the number of genes is from 1 to N and the range of the number of time points of observations is from 1 to M. Then the data available for this study was of the form $R_{ij}=\ln[x_{ij}(37)/x_{ij}(26)]$. Let the grand mean of all the observed ratios be denoted by μ

$$\mu = \frac{1}{N \cdot M} \sum_{ij} R_{ij} \tag{12}$$

The total sum of squares used to calculate the variance is given by

$$Z = \sum_{ij} \left(R_{ij} - \mu \right)^2 \tag{13}$$

Denote the log of the ratios of the modeled results by $Q_{ij}=\ln[m_{ij}(37)/m_{ij}(26)]$. Then the sum of squares of the residuals is given by

$$S = \sum_{ij} \left(R_{ij} - Q_{ij} \right)^2 \tag{14}$$

We define the fraction of the variance explained by the model as r^2

$$r^2 = 1 - \frac{S}{Z} \tag{15}$$

We use S as the cost function upon which to sort the members of the population at each generation. That is, the members preserved for the next generation have the lowest cost and hence maximum r^2.

The total sum of squares used to calculate the variance for the data in Table 2 is Z=11.393.

Tabulation of results. In Table 4 we show the results of the genetic algorithm broken out by the restrictions on the form of the regulations allowed and whether a single chromosome or multiple chromosomes were used. We also give the mutation rate used to get this result. We found that the results were highly sensitive to the mutation rate and the results reported here are for the mutation rates that produced the best cost function. The full chromosomes for the best two members of the population for each case are given in Appendix A.

We see in Table 4 that the solution that accounts for most of the variance is solution 5 for `irestrict=5`, which is the case such that the network chromosomes have no restrictions. This was found by the Multiple Chromosome approach. However, in general it appears that the Single Chromosome approach produces slightly better results on a slightly more consistent basis.

In Table 5 we show the **`prombelong`** arrays for each solution as numbered in Table 4. Recall **`prombelong`** fixes membership of the genes to the promoters. The names (numbers) of the promoters are irrelevant. The important result is the groupings of genes defined by **`prombelong`**. We see that ADU are included in the same group three time; BE is grouped together once; CT is grouped together 7 times; FK is grouped together 6 time; G is isolated by itself only twice; HP are grouped together 8 times; IM are grouped 4 times; JO is grouped once; L is isolated by itself 9 times; NS is grouped only once; Q is isolated by itself 10 times; R is isolated twice; and V is isolated 7 times.

In Table 6 we show the **`listprotatop`** array for the 11 solutions mentioned in Table 4. Any instance of violation of the restrictions given in Table 3 arrives in the solutions through mutations of parents. While there may be "selection pressure" for such violations, it is not as strong as it appears at first glance. For example in solution 1, the last two promoters apparently violate the restriction that no protein both activates and represses. However upon examination of the sign of the c_1 for the last two promoters, in both cases these promoters have two repressor operators at that location.

In Table 7, we give the signs of the c_1's in the **`k1`** array. These are very useful in interpreting the **`listprotatop`** array. We give the activity of the putative external protein at 37 degrees relative to its activity at 26 degrees. Note that most solutions give an increase of about a factor of 2; however one solution gives about a 14 fold increase in activity and another solution gives a reduction of about a half.

The full details of all 11 solutions are given in the appendix. In Fig. 2, we show the plots of the three best solutions as indicated by their r^2 and the corresponding data points. In the main the best few models do a good job fitting the data. At a few data points the models are a little weak.

Table 4. Best cost function found for each restriction type (Table 3). Associated mutation rate which produced cost and associated r^2 are shown also. All values were found using an upper bound of 100.0 on c_p except as noted. Results are broken out by Single Chromosome approach and Multiple Chromosome approach.

Solution	`irestrict`	Mutation rate	Cost	$r2$
		Multiple Chromosomes Approach		
1	1	0.004	2.349	0.79
2	2	0.008	2.695	0.76
3	3	0.005	2.279	0.80
4	4	0.007	2.895	0.75
5	5	0.008	1.907	0.83
		Single Chromosome Approach		
6	1	0.0075	2.178	0.81
7	2	0.008	2.293	0.80
8	3	0.007	2.227*	0.80
9	3	0.0075	2.350†	0.79
10	4	0.0065	2.115	0.81
11	5	0.006	2.242	0.80

*We obtained this value using an upper bound of 10.0 on c_p.
†Best solution obtained for `irestrict`=3 with an upper bound of 100.0 on c_p in the Single Chromosome approach.

Table 5. The **`prombelong`** array for the 11 best solutions (lowest cost, highest r^2) as found by the genetic algorithm. See Table 4.

Solution	**`prombelong`**																				
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	1	2	3	3	4	2	5	6	7	8	2	9	7	10	4	6	11	4	12	2	3	13
2	1	2	3	4	5	2	5	6	3	2	2	7	8	9	5	10	11	5	7	12	1	9
3	1	2	3	4	5	3	2	6	7	2	2	8	9	10	5	11	12	5	13	3	1	13
4	1	2	2	3	4	2	5	6	7	2	2	8	9	10	4	6	11	4	12	2	3	13
5	1	2	2	1	3	2	3	4	5	2	6	7	8	9	3	4	10	11	12	2	1	12
6	1	2	3	1	4	2	4	5	6	3	2	7	6	8	4	5	9	4	10	3	11	8
7	1	2	2	3	4	2	5	6	7	4	5	8	9	10	4	6	11	4	12	2	3	13
8	1	2	2	3	4	2	4	5	6	2	7	8	6	9	10	5	11	4	12	2	1	13
9	1	2	2	3	4	5	4	6	7	2	2	7	8	9	4	6	10	11	9	12	1	13
10	1	2	3	1	2	3	4	5	6	2	3	7	8	9	10	5	11	4	12	6	1	13
11	1	2	3	1	4	2	4	5	6	7	2	8	6	9	10	11	12	4	13	3	1	12

Table 6a. The **listprotatop** array for the 13 promoters for the 11 solutions. The gene numbers given in the odd locations are activator proteins if $c1$ positive, else repressors. The gene numbers in the even locations are repressor proteins.

Soln.	**listprotatop**																								
	Promoters																								
	1		2		3		4		5		6		7		8		9		10		11		12		13
	Operators																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	21	1	19	23	22	23	20	22	6	17	4	10	14	17	18	23	5	17	15	7	3	7	10	10	2	2
2	16	23	16	22	5	20	23	16	13	23	5	15	20	14	20	7	20	5	9	3	17	9	13	23	5	15
3	17	23	5	4	19	6	1	23	12	23	12	23	12	19	4	4	12	19	22	11	19	6	22	9	18	9
4	23	9	7	23	9	23	23	17	11	23	11	2	3	21	13	21	17	21	20	9	22	9	13	20	14	20
5	22	22	9	23	11	19	16	14	12	21	14	19	18	18	12	14	19	13	14	13	23	19	23	9	6	20
6	7	19	4	22	11	23	9	23	4	22	12	4	17	4	1	18	14	13	16	18	8	23	2	4	20	17
7	13	13	5	1	23	18	12	23	5	1	5	1	20	1	1	1	19	10	11	21	17	11	6	21	1	5
8	23	17	11	16	18	23	12	23	2	1	11	15	12	23	9	15	22	13	12	23	22	13	11	7	4	7
9	19	23	12	23	19	23	12	23	12	23	20	4	20	18	12	19	17	2	22	9	13	23	18	23	15	9
10	12	23	15	23	2	15	2	22	1	22	2	15	10	17	17	15	19	20	2	23	14	20	20	10	14	10
11	5	12	14	22	15	22	10	22	13	14	10	15	23	15	23	14	9	18	23	22	6	16	7	13	14	18

Table 6b. This table is the same as Table 6a, except it is rewritten to clarify the proteins , which control the operators. As before, the proteins at the odd operators are activator proteins if $c1$ positive, else repressors. The proteins in the even locations are repressor proteins. The external protein 23 is designated by X.

Soln.	**listprotatop**																								
	Promoters																								
	1		2		3		4		5		6		7		8		9		10		11		12		13
	Operators																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	U	A	S	X	V	X	T	V	F	Q	D	J	N	Q	R	X	E	Q	O	G	C	G	J	J	B	B
2	P	X	P	V	E	T	X	P	M	X	E	O	T	N	T	G	T	E	I	C	Q	I	M	X	E	O
3	Q	X	E	D	S	F	A	X	L	X	L	X	L	S	D	D	L	S	V	K	S	F	V	I	R	I
4	X	I	G	X	I	X	X	Q	K	X	K	B	C	U	M	U	Q	U	T	I	V	I	M	T	N	T
5	V	V	I	X	K	S	P	N	L	U	N	S	R	R	L	N	S	M	N	M	X	S	X	I	F	T
6	G	S	D	V	K	X	I	X	D	V	L	D	Q	D	A	R	N	M	P	R	H	X	B	D	T	Q
7	M	M	E	A	X	R	L	X	E	A	E	A	T	A	A	A	S	J	K	U	Q	K	F	U	A	E
8	X	Q	K	P	R	X	L	X	B	A	K	O	L	X	I	O	V	M	L	X	V	M	K	G	D	G
9	S	X	L	X	S	X	L	X	L	X	T	D	T	R	L	S	Q	B	V	I	M	X	R	X	O	I
10	L	X	O	X	B	O	B	V	A	V	B	O	J	Q	Q	O	S	T	B	X	N	T	T	J	N	J
11	E	L	N	V	O	V	J	V	M	N	J	O	X	O	X	N	I	R	X	V	F	P	G	M	N	R

Table 7. The signs of the c_1's in the $\mathbf{k}_1$ array. If c_1 is negative for that promoter, then the odd operator given in Table 6 is a repressor operator. If c_1 is positive then the odd operator for that promoter given in Table 6 is an operator for activation. We also include the activity of the external protein at 37 degrees relative to its activity at 26 degrees.

Solution	Signs of c_1 in $\mathbf{k}_1$ array													Activity external protein
	1	2	3	4	5	6	7	8	9	10	11	12	13	
1	+	+	−	+	+	+	+	+	−	−	+	−	−	2.1117
2	−	+	+	−	+	+	−	−	+	+	+	−	+	2.9511
3	−	+	−	+	+	+	+	−	+	+	−	+	−	3.6087
4	−	−	−	−	+	+	+	+	−	−	+	+	+	5.9769
5	−	+	+	−	+	−	−	+	+	+	−	+	+	1.8267
6	+	−	+	+	−	+	−	+	+	+	+	−	+	2.3496
7	−	+	−	+	+	+	+	−	+	−	+	+	+	2.6066
8	−	+	+	+	+	+	+	+	+	+	+	+	+	9.4696
9	−	+	−	+	+	+	+	+	+	+	+	−	−	2.6897
10	+	−	+	+	−	+	−	−	+	+	+	−	+	14.5612
11	+	+	+	+	+	+	+	−	−	+	+	−	−	0.4889

DISCUSSION

The genetic algorithm does indeed find apparent sets of equations, which produce results that do a remarkable job at reproducing the data. However several critical observations are in order.

First, it might be incorrect to refer to these results as solutions, meaning that extema have been found. Note that the several values of **x2max** crowd the upper bound of the range available to it. It may be the case that there are no true extrema. We have 66 data points and 62 real parameters and 47 integer parameters.

Secondly it is evident that while many promising patterns emerge from the results found by the GA, it is also clear that many solutions are found which produce very similar results. They obviously can't all be the correct description of the true dynamics of the genetic network. A classic problem with genetic algorithms is that there is no guarantee that they will converge to the global maximum.

However, what the GA does provide is a group of possible solutions, which can be falsified with more experiments. In particular "knocked-out gene" experiments offer a good possibility to make progress. The models found initially by the GA can be adapted to particular knock-out experiments by setting appropriate c_p 's for the appropriate gene expression equations to zero. Resulting runs of the models can be compared to experimental results. In this way some models can be rejected and others provisionally accepted. As more information is gathered in this way, the GA can be altered to generate new models in case the true global optimum was not found in the first round. That is, if by new experiments, one can exclude certain possibilities or fix some portion of the network the GA can be modified to honor this information. For example, if one determined by some independent means, such as further experimentation, that two genes must belong to the same promoter, one could generate all initial chromosomes with this property and exclude mutation in that region of the chromosome. Thus only solutions with the property in question could be found.

In short we suggest that the real power of this approach is in the iterations possible between experiment and models that could ultimately determine the interactions in genetic regulation.

Once the correct equations are found, a state-based description could be used to encode the information contained in the ODE's. Such an encoding could be used in higher-level models in which it is unnecessary to simulate the details of gene-regulation, but in which the results of gene regulation are critical.

Acknowledgments

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

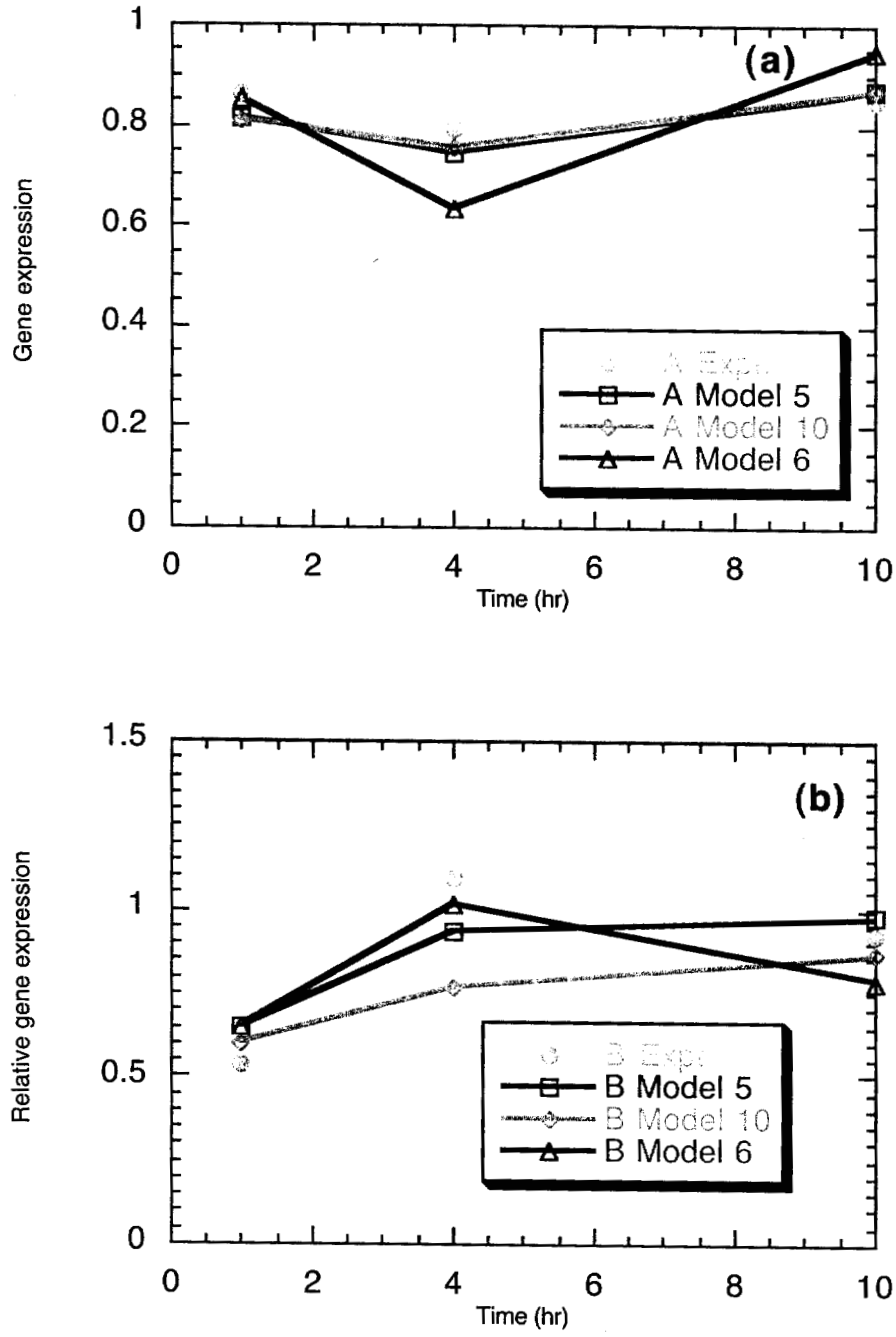


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (a) gene A and (b) gene B.

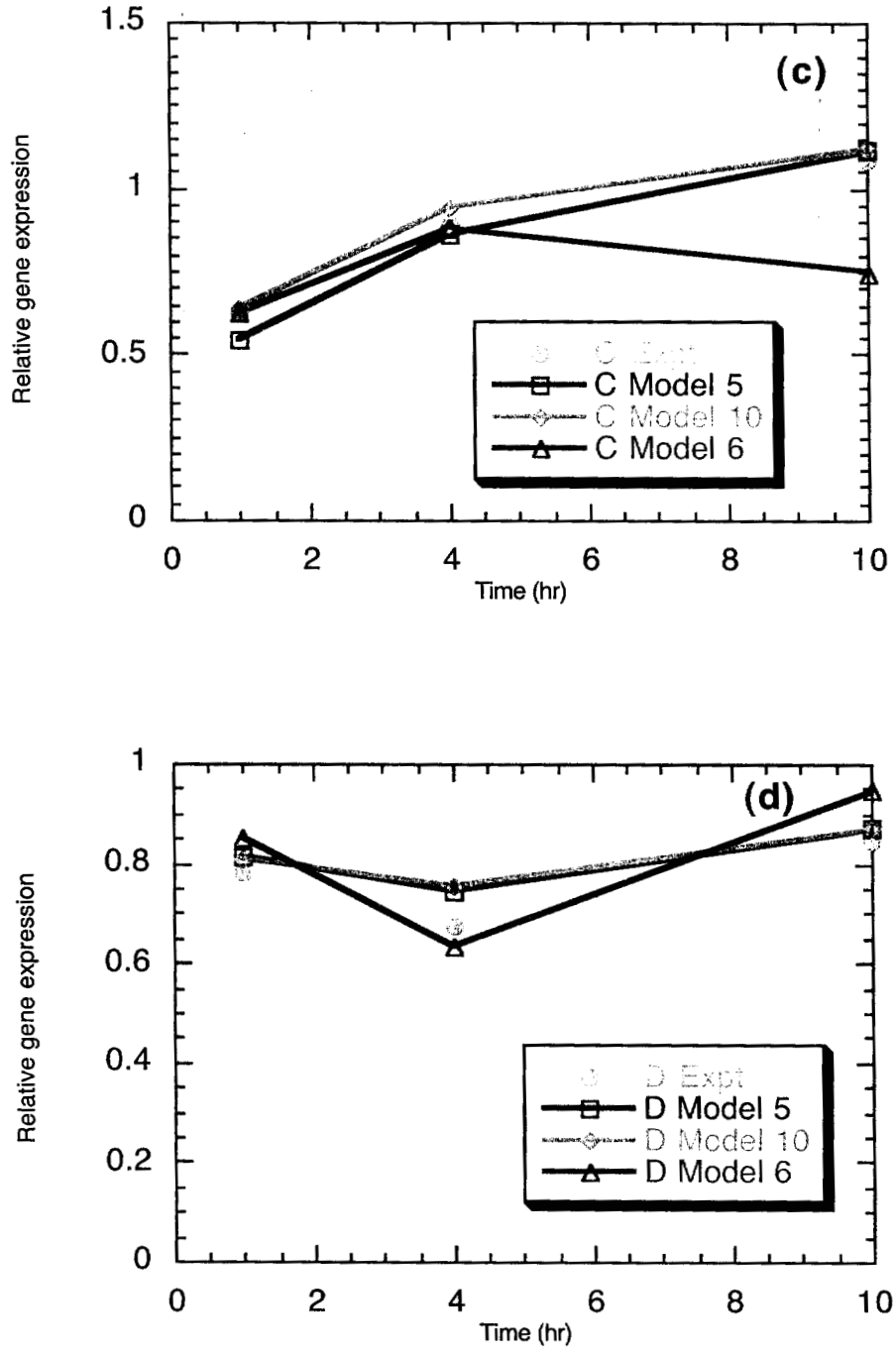


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (c) gene C and (d) gene D.

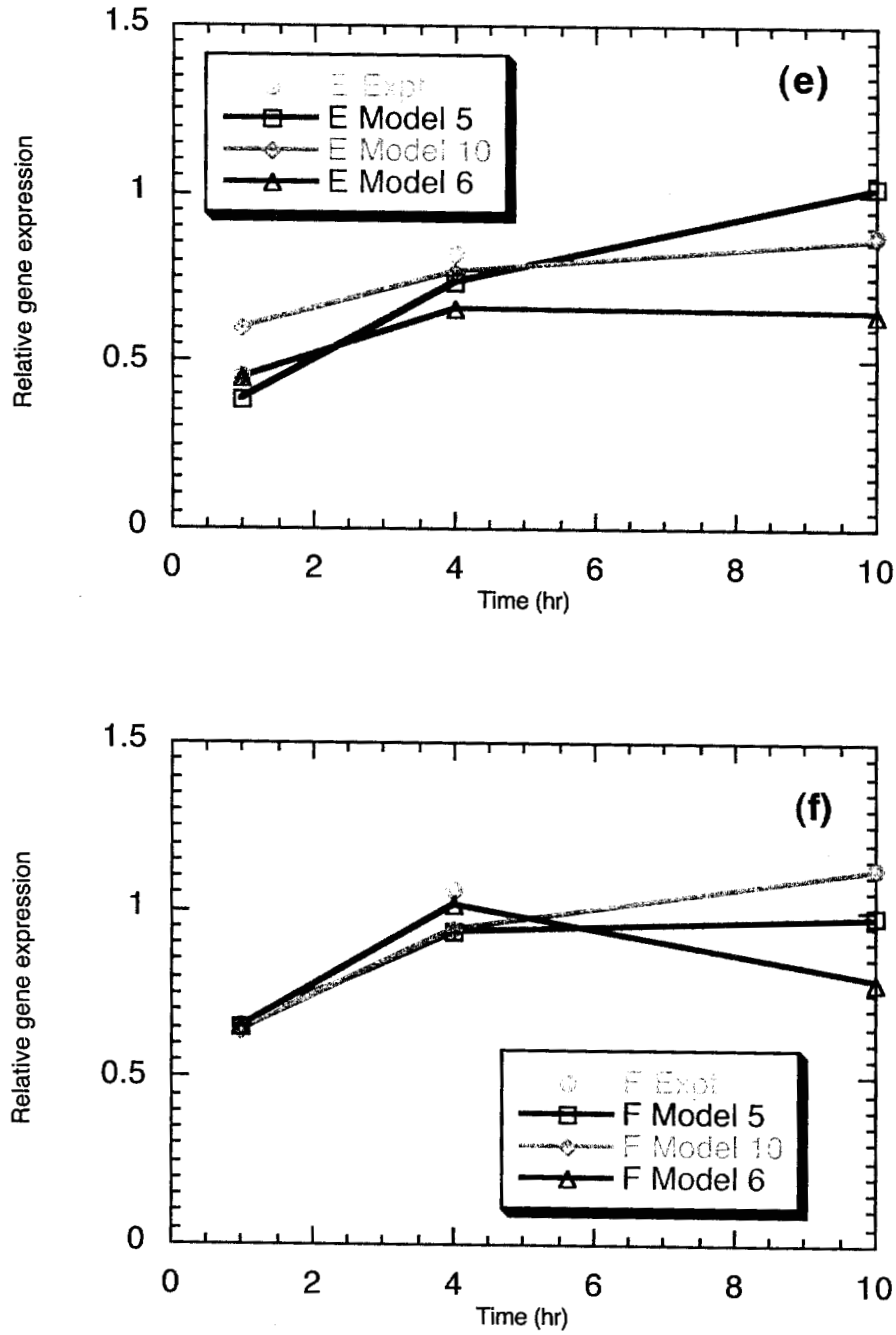


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (e) gene E and (f) gene F.

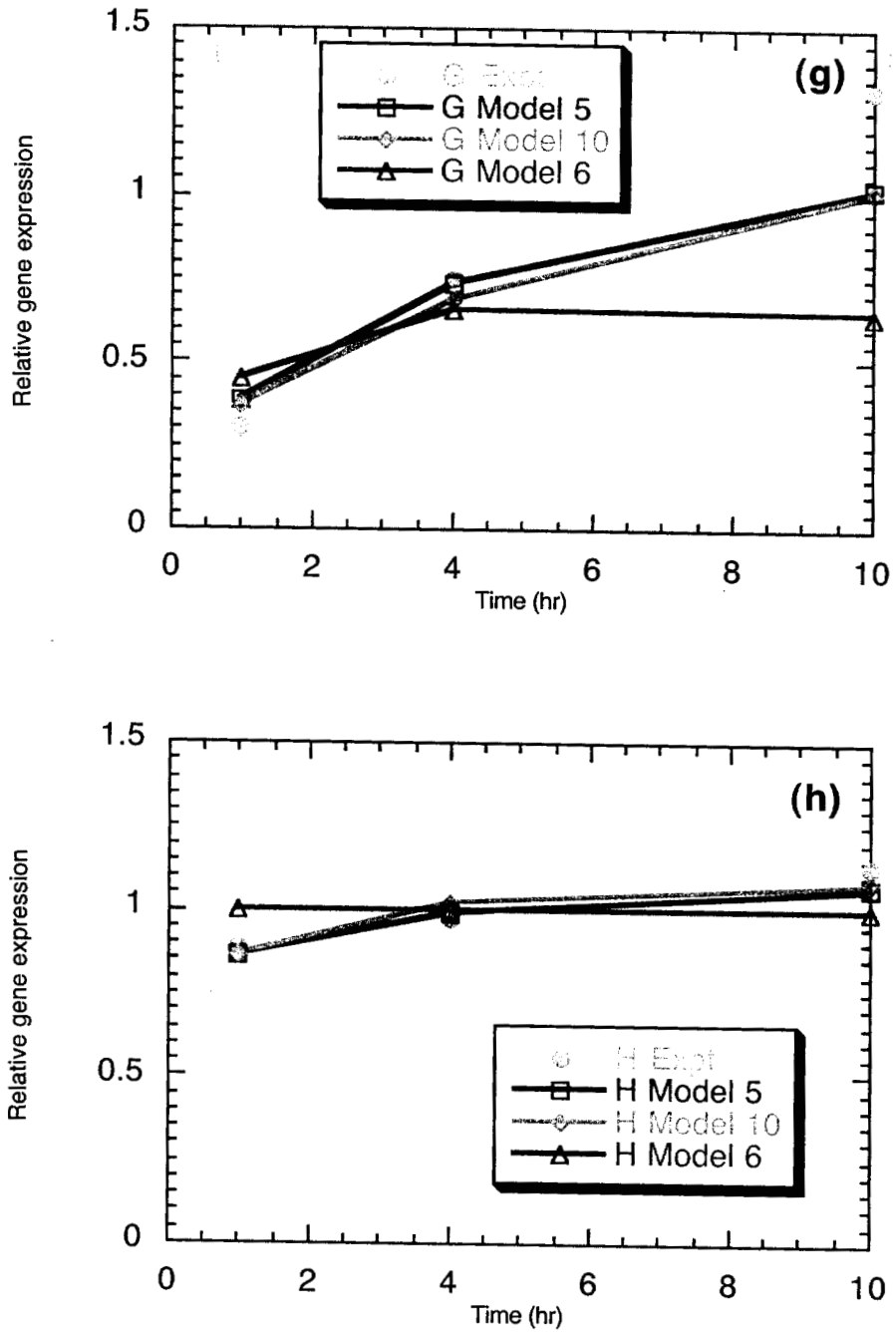


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (g) gene G and (h) gene H.

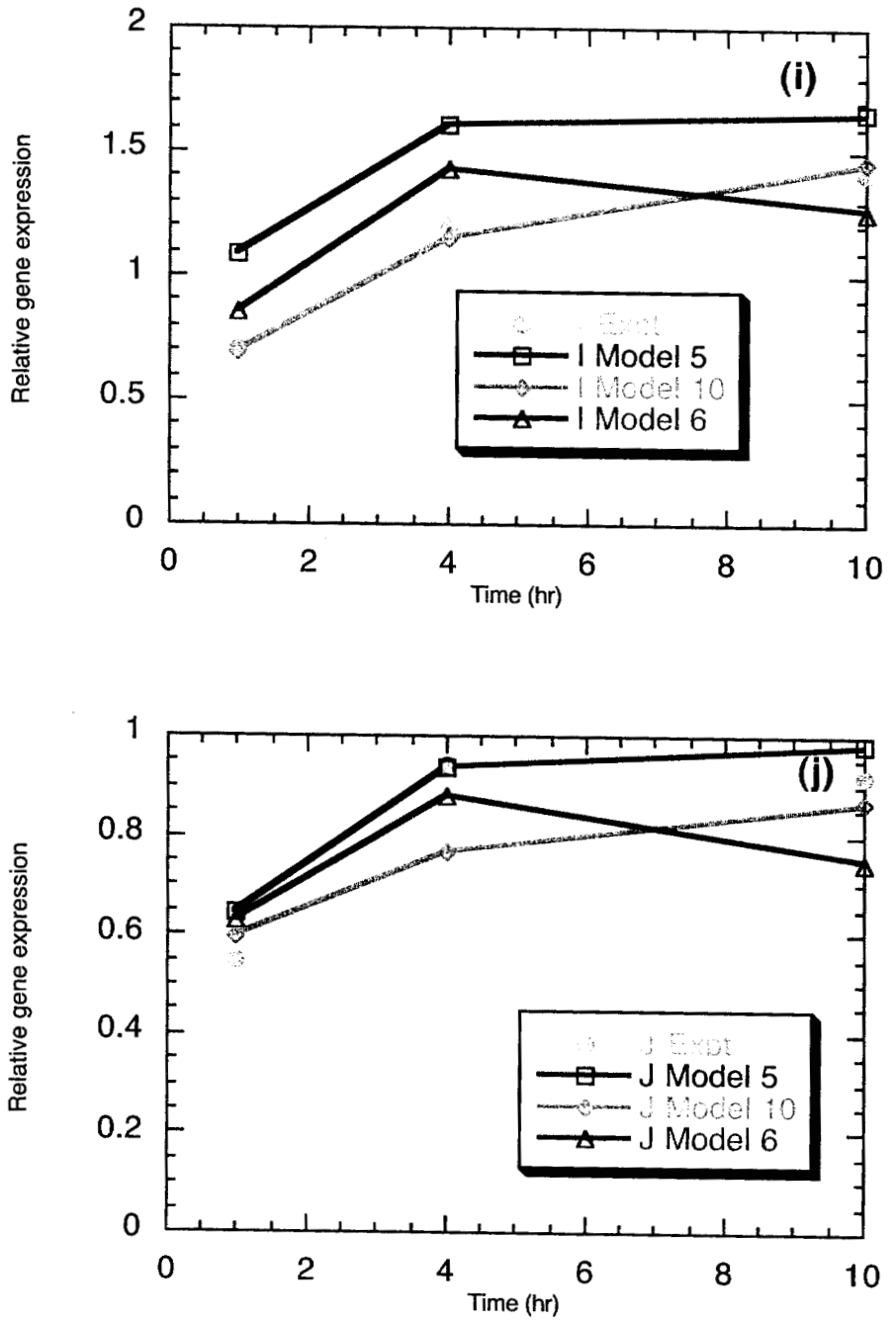


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (i) gene I and (j) gene J.

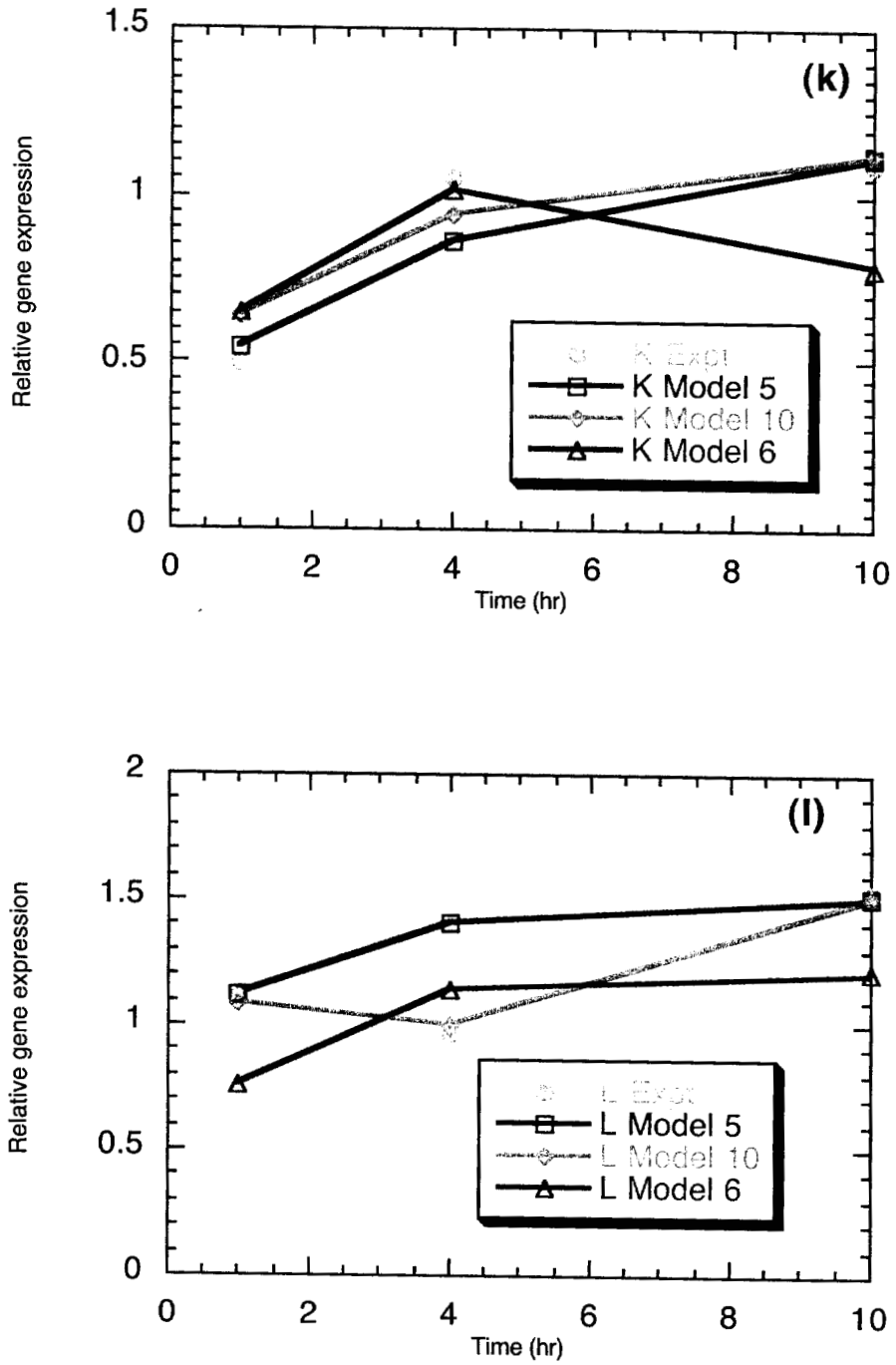


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (k) gene K and (l) gene L.

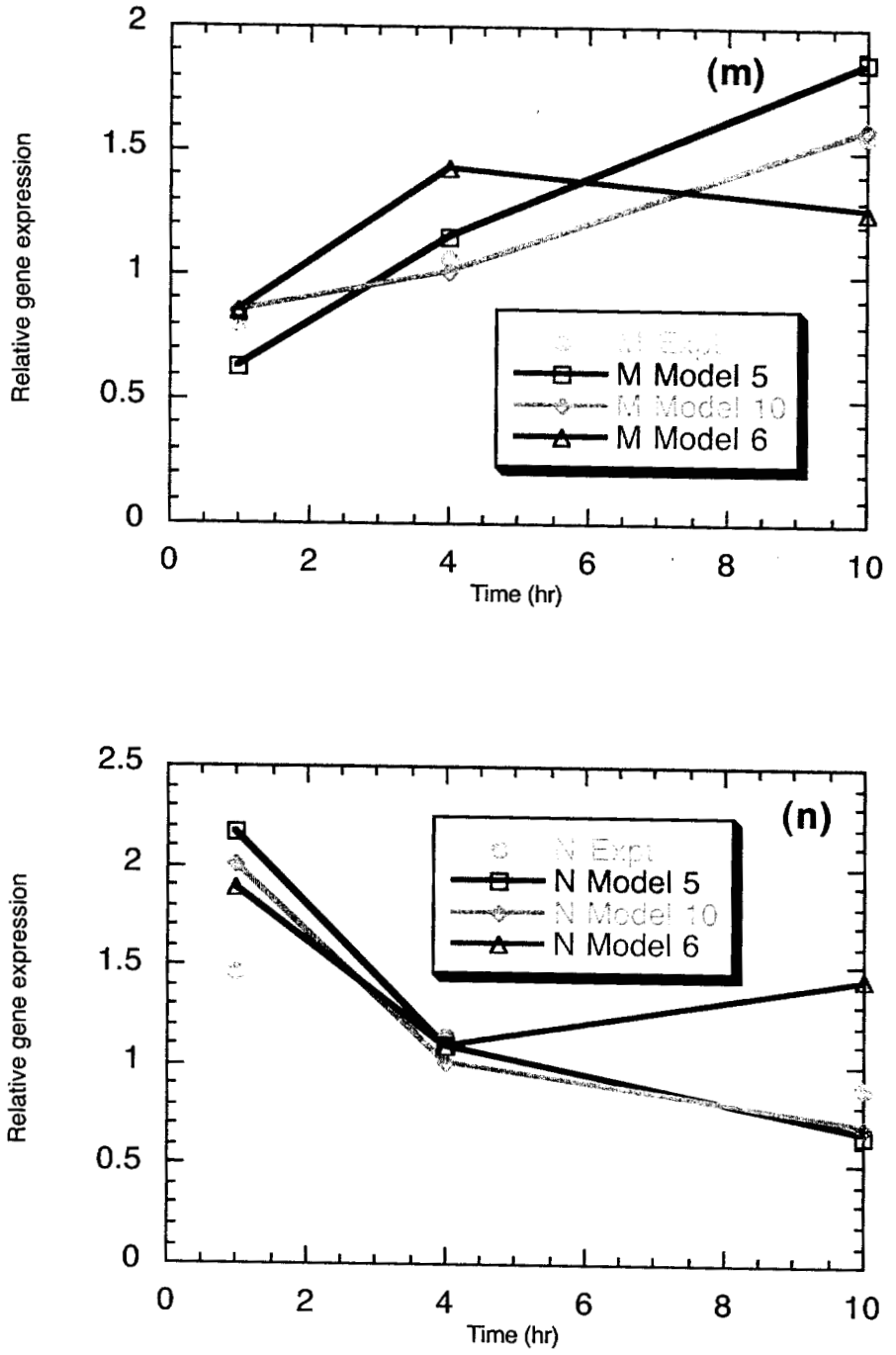


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (m) gene M and (n) gene N.

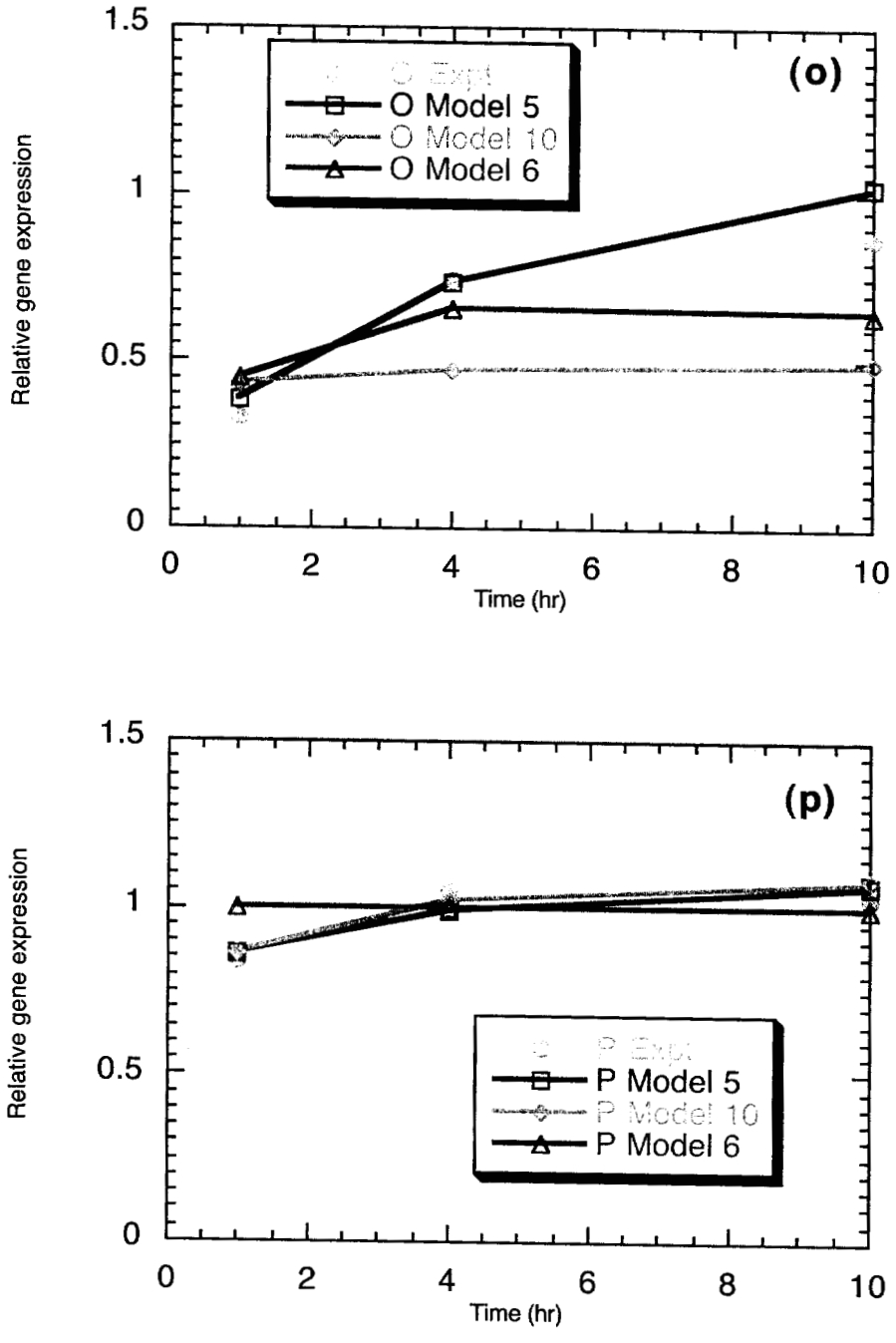


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (o) gene O and (p) gene P.

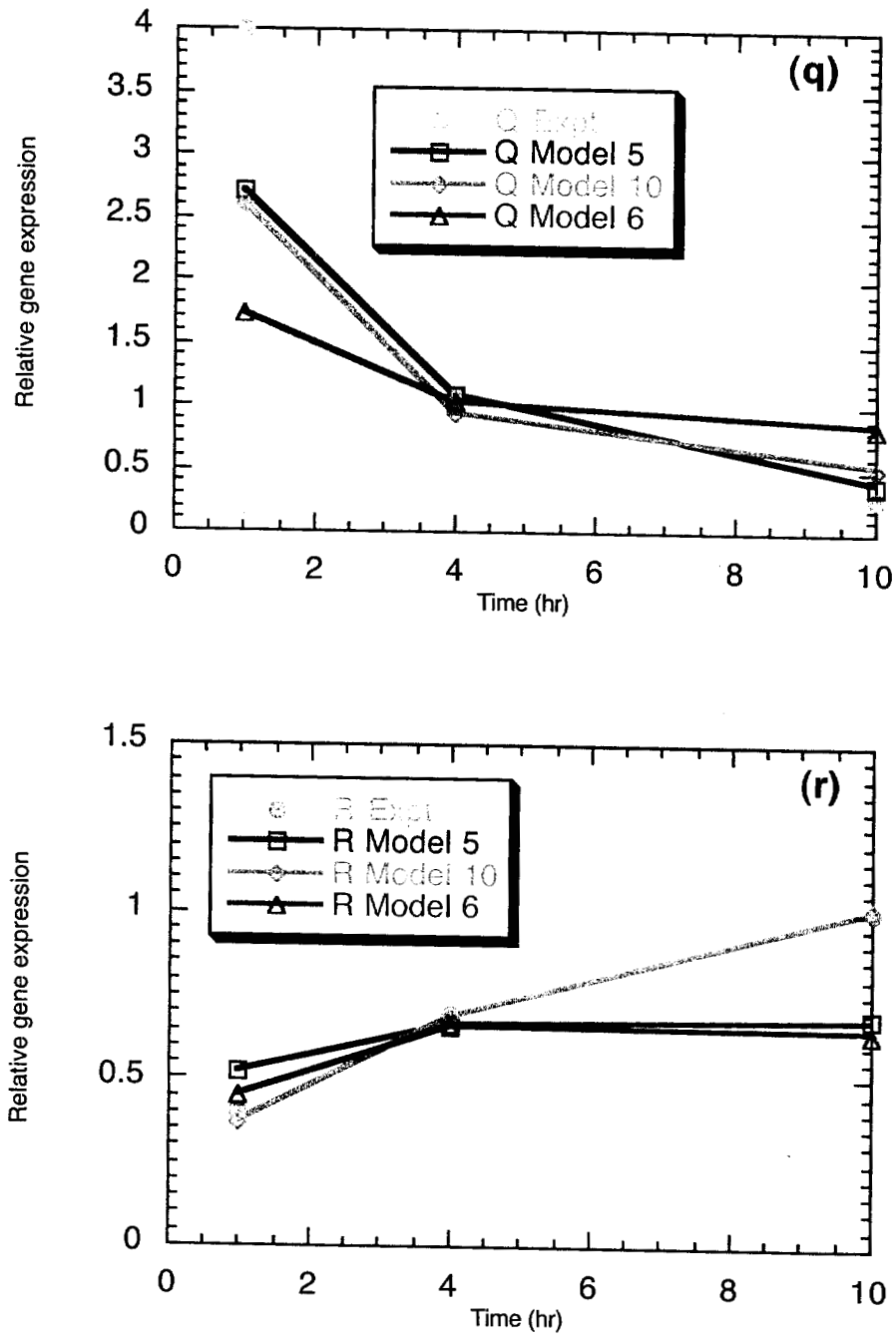


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (q) gene Q and (r) gene R.

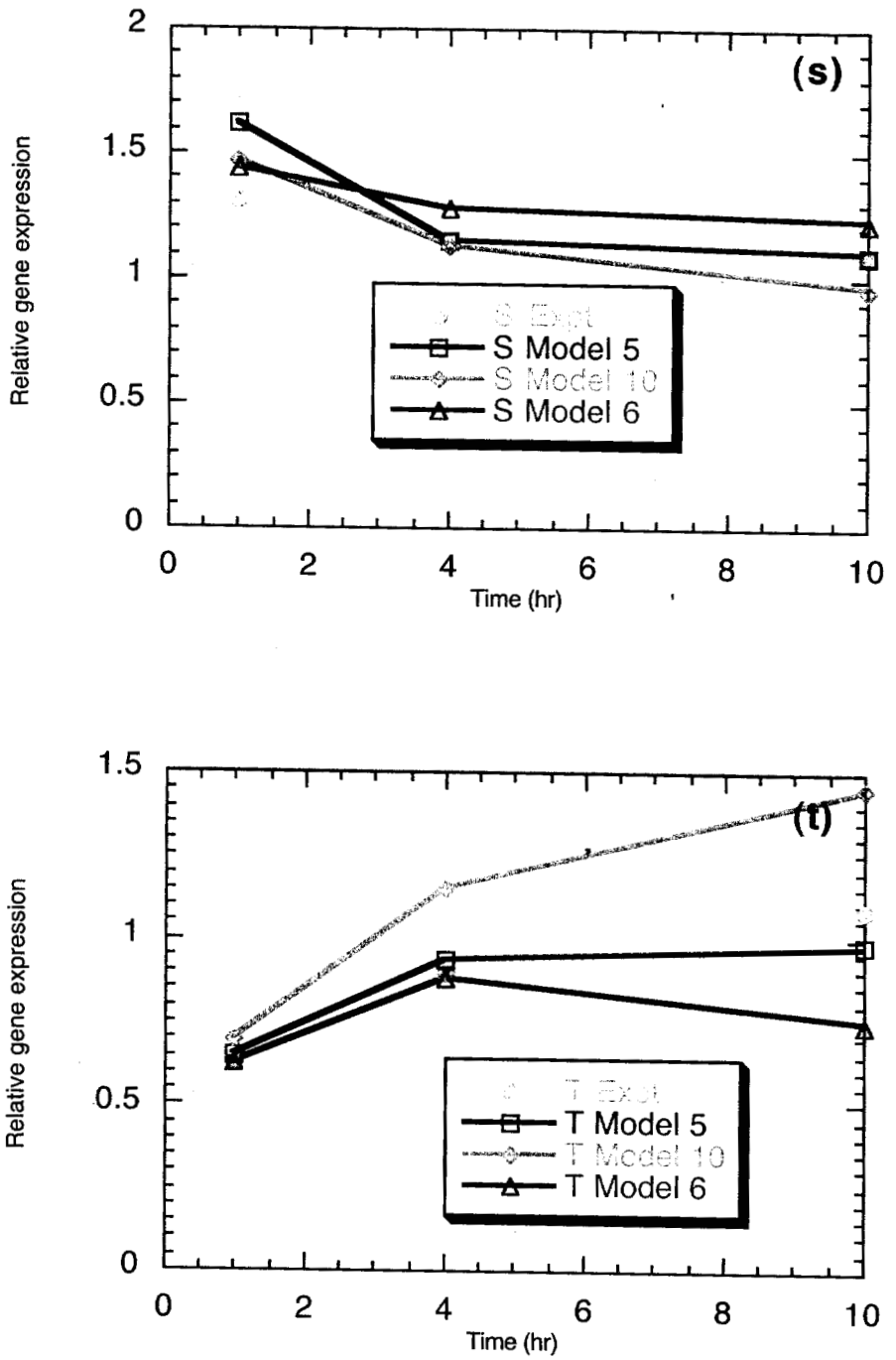


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (s) gene S and (t) gene T.

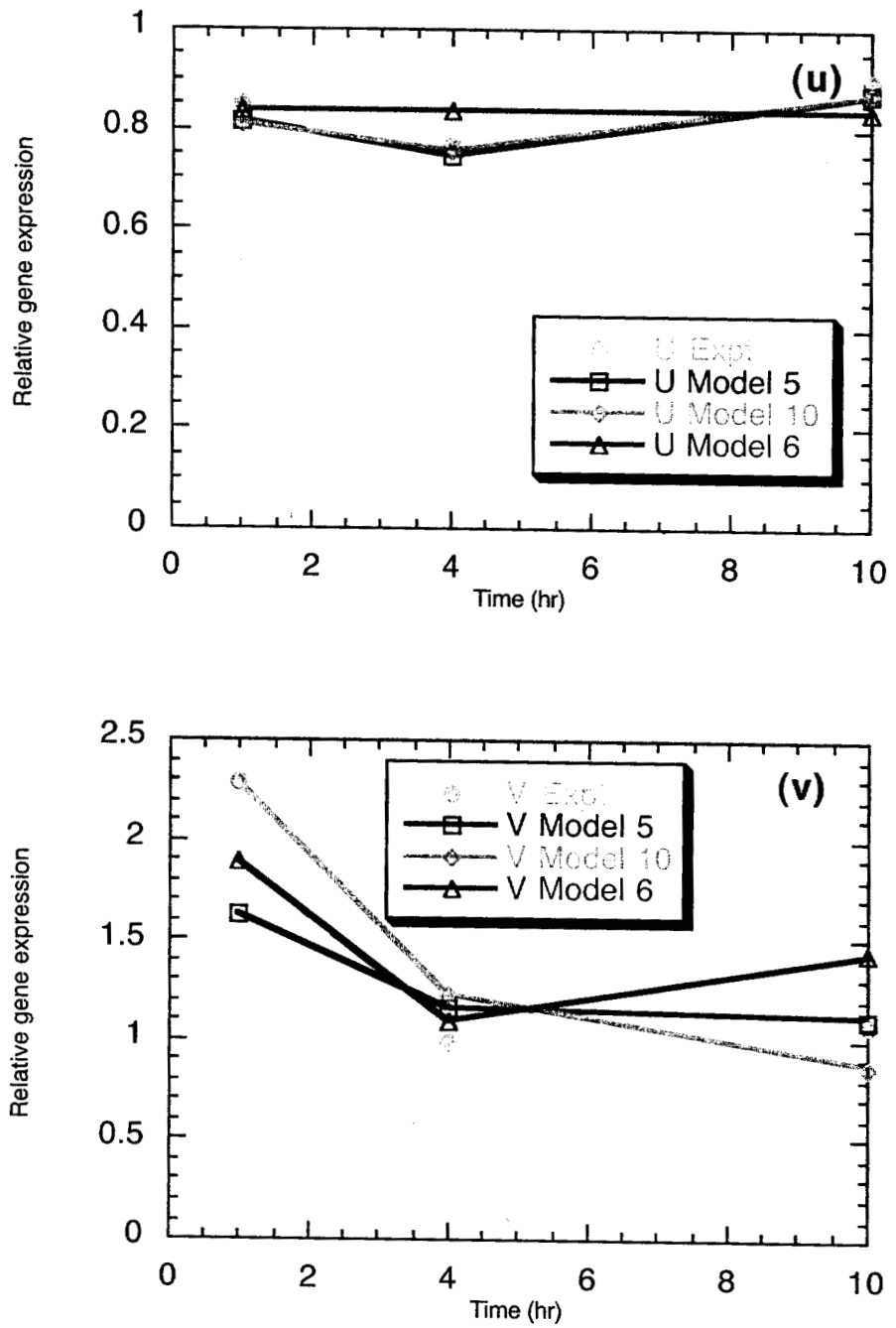


Fig. 2. Comparison of the results of the three models with the highest r^2 to the expression data for (u) gene U and (v) gene V.

REFERENCES

Fogel, G.B., D.W. Corne. 2003. Evolutionary computation in bioinformatics. Morgan Kaufmann Publ.

Gibson, M.A., E. Mjolsness. 2001. Modeling the activity of single genes. p. 1-48. In Computational Modeling of Genetic and Biochemical Networks. (J.M. Brower, H. Bolouri, eds.) MIT Press.

Goldberg, D.E. 1989. Genetic algorithms. Addison-Wesley.

Goldberg, D.E. 2002. The design of innovation. Kluwer Academic.

Haupt, R.L., S.E. Haupt. 1998. Practical genetic algorithms. John Wiley & Sons.

Holland, J.H. 1962. Outline for a logical theory of adaptive systems. Journal of the Association for Computing Machinery 3:297-314.

Holland, J.H. 1973. Genetic algorithms and the optimal allocation of trials. SIAM Journal of Computing 2(2):88-105.

Holland, J.H. 1975. Adaptation in natural and artificial systems. University of Michigan Press.

Johnson V.M., L.L. Rogers. 2001. Applying soft computing methods to improve the computational tractability of a subsurface simulation-optimization problem. J. PETROL. SCI. ENG. 29 (3-4): 153-175.

Kercher, J.R. 2003. State-based automata descriptions of cellular gene regulation and protein kinetics. LLNL UCRL-ID-151868

Koza, J.R., W. Mydlowec, G. Lanza, J. Yu, M. A. Keane. 2000. Reverse engineering and automatic synthesis of metabolic pathways from observed data using genetic programming. Stanford University Technical Report SMI-2000-0851.

Mitchell, M. 1996. An introduction to genetic algorithms. MIT Press.

Reeves, C.R. 1993. Genetic algorithms. p. 151-196. In Modern Heuristic Techniques for Combinatorial Problems (C.R. Reeves, ed.) John Wiley & Sons.

Rogers L.L., F.U. Dowla, V.M. Johnson. 1995. Optimal field-scale groundwater remediation using neural networks and the genetic algorithm. ENVIRON. SCI. TECHNOL. 29:1145-1155.

Somogyi, R., S. Fuhrman, X-L. Wen. 2001. Genetic network inference in computational models and applications to large-scale gene expression data. p. 119-160. In Computational Modeling of Genetic and Biochemical Networks. (J.M. Brower, H. Bolouri, eds.) MIT Press.

Thomas, R., R. D'Ari. 1990. Biological feedback. CRC Press.

Voit, E.O. 2000. Computational analysis of biochemical systems. Cambridge University Press.

Appendix

The following are the detailed results of the 11 solutions found by the genetic algorithm. These solutions are referenced in the text in Table 4. In each case the genetic algorithm code printed out the two best solutions found at the end of the run.

Solution 1.

```
generation=        10800
imember=    1  k1=  3.6697E-002  1.0016E-002  -0.40743  1.0022E-002
1.0039E-002  1.3267E-002  1.0233E-002  0.41129  -1.0855  -0.15029
1.0002E-002  -0.57609  -0.29920

imember=    1  k1'=  3.9821E-002  2.8010  3.3703  8.1945  0.84960
0.71181  0.36806  9.9529  0.63750  2.5089  0.93491  0.64774  0.42153

imember=    1  x2max=   10.357  60.901  12.559  99.879  70.207  74.799
28.615  15.645  55.878  26.072  64.425  99.248  99.865  2.1116

imember=    1  k2=  7.9113E-002   4.7153  0.20904  2.1871  4.9820  4.9952
4.9903  1.9917  1.0184  0.47868  0.30230  0.40430  0.89654  0.82771
2.8338  0.66469  4.9984  4.9974  1.6760  4.9857  0.96324  1.5567

imember=    1  listprotatop=  21  1  19  23  22  23  20  22  6  17  4  10
14  17  18  23  5  17  15  7  3  7  10  10  2  2

imember=    1  prombelong=  1  2  3  3  4  2  5  6  7  8  2  9  7  10  4
6  11  4  12  2  3  13


imember=    2  k1=  3.6553E-002  1.0016E-002  -0.40743  1.0022E-002
1.0039E-002  1.3599E-002  1.0233E-002  0.41129  -1.0855  -0.15029
1.0002E-002  -0.57609   -0.29920

imember=    2  k1'=  3.9854E-002   2.8010  3.3703  8.1945  0.84958
0.71433  0.36825  9.9529  0.63750  2.5089  0.93491  0.64784  0.42153

imember=    2  x2max=   10.364  60.901  12.559  99.879  70.207  72.834
28.761  15.645  55.878  26.072  64.425  99.248  99.865  2.1116

imember=    2  k2=  7.5943E-002   4.7153  0.20904  2.1873  4.9820  4.9952
4.9903  0.72854  0.69500  0.47868  0.63573  1.6017  0.68416  0.82114
2.8338  1.1031  4.9984  4.9974  1.6759  4.9857  0.96322  1.5567

imember=    2  listprotatop=  21  1  19  23  22  23  20  22  6  17  4  10
14  17  18  23  5  17  15  7   3  7  10  10  2  2

imember=    2  prombelong=  1  2  3  3  4  2  5  6  7  8  2  9  7  10  4
6  11  4  12  2  3  13
   costmin=   2.3487 costmean=   4.3249  coststddev=   12.056  end
generation=        10800
```

Solution 2

```
generation=        11750
imember=    1  k1= -0.11400  1.0343E-002  2.4763E-002 -0.44840  1.0041E-
002  5.9741E-002 -0.16311  -0.74530  1.0062E-002  8.1900E-002  1.0062E-
002  -0.62432  5.9741
imember=    1  k1'=  0.12381  0.53694  2.9520  0.88175  1.7221  9.3478E-
002  2.6270E-002  1.2201E-002  2.0471  0.19986  1.7822  0.41384  9.3478
imember=    1  x2max=   34.601  95.913  90.977  13.925  72.472  10.440
85.508  99.676  99.978  68.565  99.983  12.337  10.440  2.9511

imember=    1  k2=  0.70473  1.4632  0.57315  4.0772  4.9632  0.23188
0.26745  6.8968E-002  4.8357  0.11296  0.28879  0.35478  1.3332  0.10234
0.36756  4.9214  4.6095  0.85651  0.33105  0.47778  0.54377  4.9877

imember=    1  listprotatop=  16  23  16  22   5  20 23  16  13  23   5
15 20  14  20   7  20   5   9   3  17   9  13  23   5  15

imember=    1  prombelong=   1   2   3   4   5   2   5   6   3   2   2
7   8   9   5   10  11   5   7  12   1   9

imember=    2  k1= -0.27026  1.3002  1.8818E-002 -0.32426  1.0058 -
0.14131 -0.73773  1.2014E-002   8.6262E-002  1.0092E-002   1.9278 -
0.61402  -1.5804
imember=    2  k1'=  0.12874  0.44003   2.9928  0.77031   1.7369
2.5463E-002  1.0446E-002   2.0171  0.30988   1.7544   1.6103  0.42072
1.4673
imember=    2  x2max=   45.325   66.697   93.830   15.584   87.782
56.422   99.659   97.719   66.131   99.226   6.1271   13.806   25.724
2.9644
imember=    2  k2=  0.13705  0.24452  0.60667  0.78171   4.9638  8.4974E-
002  0.25618  0.69104   4.8797  0.49112   1.2146   3.5867   1.3602
0.10008  0.98113   4.7324  0.55298   1.0640   1.3024  0.48105  0.74245
4.9284
imember=    2  listprotatop=  16  23  16   22   5  20  23  16  13   23
20  14  20   7  20    5   9   3  22   9  10    5  13  23   8   5

imember=    2  prombelong=   1   2   3    4   5   2   5   6   3    2   2
6   7   8   5    9  10   5  11  12   1
    8
   costmin=   2.6948 costmean=   5.5560  coststddev=   7.8041  end
generation=        11750
```

Solution 3

```
generation=        11750
imember=    1  k1= -1.1314E-002  1.0003  -1.7609  1.0094E-002  1.0004
0.21651  1.0110E-002 -0.22920  1.1259E-002  0.48571
-0.48690  1.0015E-002 -0.41776
imember=    1  k1'=  0.16430  0.14333  4.5860E-002  0.15254  0.91452
0.10096   4.0675  0.29303   6.8063  1.0694E-002  0.81579   2.1554
0.22047
imember=    1  x2max=   3.6681   99.818   99.957   2.0768   34.967
2.6535   93.413   99.480   6.7040   12.478   74.252   99.955   99.955
3.6086
```

```
imember=    1  k2=  0.58115    2.5330   0.20394   0.56259    4.9986   0.37572
0.17290   0.26642    4.9993   0.22230    4.9583    1.5551   0.20882   0.69974
0.18917    2.4194   0.54359    4.9993    4.9981   0.18820    2.8670    4.9713
imember=    1  listprotatop=  17  23   5    4  19   6   1  23  12   23
12  23  12  19   4    4  12  19  22  11  19    6  22   9  18   9

imember=    1  prombelong=   1   2   3   4   5   3   2   6   7   2   2
8   9  10   5   11  12   5  13   3   1   13

imember=    1  prottypeact= T F F F T F F T F F F T F F T T
F F F F T T  F

imember=    2  k1= -1.13149E-002  1.0003  -1.7609  1.0094E-002  1.0004
0.21651  1.0110E-002 -0.22920  1.1259E-002  0.48571
-0.48690  1.0015E-002 -0.41776
imember=    2  k1'=  0.16430  0.14333  4.5860E-002  0.15254  0.91452
0.10096    4.0675   0.29303    6.8063  1.0694E-002  0.81213    2.1554
0.22047
imember=    2  x2max=    3.6681    99.818    99.957    2.0768    34.967
2.6530    93.411    99.480    6.7040    12.877    74.252    99.955    99.955
3.6086
imember=    2  k2=  0.58115    2.3812   0.19585   0.56259    4.9986   0.37572
0.17290   0.26642    4.9993   0.22230    4.9583    1.5551   0.20882   0.69974
0.18917    1.4279   0.54359    4.9993    4.9981   0.18820    2.8670    4.9713
imember=    2  listprotatop=  17  23   5    4  19   6   1  23  12   23
12  23  12  19   4    4  12  19  22  11  19    6  22   9  18   9

imember=    2  prombelong=   1   2   3   4   5   3   2   6   7   2   2
8   9  10   5   11  12   5  13   3   1   13

imember=    2  prottypeact= T F F F T F F T F F F T F F T F F F F T T F
   costmin=    2.2791 costmean=    4.8344  coststddev=    8.5988  end
generation=        11750


Solution 4

generation=        11750
imember=    1  k1=  -2.1436 -1.9443E-002  -0.15841 -0.10783  1.1018E-002
0.26941  1.0461E-002  1.4689E-002   -1.7944
-1.4162  1.0361E-002    0.70719  1.0647
imember=    1  k1'=  2.0409E-002  0.19700  0.13416    2.9304  0.21747
0.23335  0.73710  5.9874  0.57953  0.73091    3.2017    3.8906    6.0614
imember=    1  x2max=  0.79578    99.346    6.6280    96.733    83.232
38.853    98.710    9.8187    64.233    99.552    99.449    61.110    90.493
5.9768
imember=    1  k2=  0.74352  0.14980    4.9502  0.66052  0.30582  0.67115
0.36199  0.74526    4.9085  0.67343    3.1090    2.3710  0.37069    4.7033
0.13472  9.2884E-002    4.9766  0.27605  0.65800    4.9725  0.40730
4.9184
imember=    1  listprotatop=  23   9   7
   23   9  23  23  17  11
   23  11   2   3  21  13
   21  17  21  20   9  22
    9  13  20  14  20

imember=    1  prombelong=   1   2   2  3   4   2   5   6   7   2   2
8   9  10   4    6  11   4  12   2   3   13
```

```
imember=    1  prottypeact= F F T F T T F T F F T T T T T F F F F F F T F

imember=    2  k1=  -2.1436 -1.9443 -0.15841 -0.10783  1.1018  0.26941
1.0461E-002  1.4689  -1.7944  -1.4162  1.0388  0.71220  1.0647
imember=    2  k1'=  2.0409E-002  0.19700  0.13416   2.9304  0.21747
0.23335  0.73710  5.9874  0.57953  0.73047   3.2017   3.8989   6.0606
imember=    2  x2max=  0.79578   99.346   6.6280   96.733   83.232
38.853   98.710   9.8187   64.233   99.552   99.449   61.110   90.487
5.9768
imember=    2  k2=  0.74352  0.14980   4.9502  0.66052  0.30582  0.67115
0.36199  0.74526   4.9085  0.67343   3.1090   2.3710  0.37069   4.7033
0.13472  9.2884E-002   4.9766  0.56514   1.4271   4.9726  0.40730
4.2428
imember=    2  listprotatop=  23   9   7   23   9  23  23  17  11   23
11   2   3  21  13   21  17  21  20   9  22   9  13  20  14  20

imember=    2  prombelong=   1   2   2   3   4   2   5   6   7   2   2
8   9  10   4   6  11   4  12   2   3  13

imember=    2  prottypeact= F F T F T T F T F F T T T T T F
F F F F F T  F
   costmin=   2.8949 costmean=   6.6092  coststddev=   12.786  end
generation=       11750


Solution 5

generation=       11750
imember=    1  k1= -5.8843E-002  1.0302  1.0195E-002  -1.3302  1.0344 -
0.14141 -0.10197  2.2302  1.2381E-002  1.0089E-002
-0.96643  2.9753E-002   2.0110
imember=    1  k1'=  0.10928   3.6005   5.3482  0.18998  0.51007  0.61138
0.10263  0.38183  0.94683   1.9899   1.4412   2.9062   1.7017
imember=    1  x2max=   97.066   93.040   96.736   10.876   42.813
91.417   96.281   23.435   98.236   96.618   95.847   96.525  0.98845
1.8267
imember=    1  k2=  0.96101   1.5615  0.65137   1.8467  0.96878   1.9681
0.76346   1.0749   4.2623   1.2756   4.9415  0.85900   4.1242   4.9788
6.4871E-002  0.37187   1.4270  0.23831   4.9824  0.40617  0.64253
0.31067
imember=    1  listprotatop=  22  22   9   23  11  19  16  14  12   21
14  19  18  18  12   14  19  13  14  13  23   19  23   9   6  20

imember=    1  prombelong=   1   2   2   1   3   2   3   4   5   2   6
7   8   9   3   4  10  11  12   2   1  12

imember=    2  k1= -5.8876E-002  1.0451  1.0196E-002  -1.3375  1.0344 -
0.14107 -0.10200  2.2259  1.2485E-002  1.0089E-002 -0.96643  2.9752E-002
-8.5529
imember=    2  k1'=  0.10907   3.5944   5.3489  0.18926  0.50988  0.59948
0.10256  0.38183  0.94682   1.9567   1.4411   2.8945   1.0384
imember=    2  x2max=   97.070   92.991   96.736   10.826   43.117
91.084   96.420   23.434   98.236   96.618   95.847   96.525   1.9836
1.8267
```

```
imember=   2  k2=  0.94451  0.84119   1.3451  0.76921   1.4675   1.9860
0.15592   1.2508   4.3040  0.74624   4.9415  0.85900   4.0703   4.9788
0.83191  0.37005   2.5886  0.23831   4.9822  0.22763  0.64253  0.31067
imember=   2  listprotatop=  22  22   9   23  11  19  16  14  12   21
14  19  18  18  12   14  19  13  14  13  23   19  23   9   3   9

imember=   2  prombelong=   1   2   2   1   3   2   3   4   5   2   6
7   8   9   3   4  10  11  12   2   1  12
   costmin=   1.9070 costmean=   7.6448  coststddev=   28.000  end
generation=       11750


Solution 6

generation=       11550
imember=   1  k1=  1.0218E-002 -5.8965  1.0457E-002  1.0204E-002 -
0.37717  1.0378E-002 -0.18220  1.0418  1.0001E-002  1.6434E-002  1.4257
-2.5969  2.4823
imember=   1  k1'=  0.19810  0.69386  7.8144E-002   4.0744  0.64698
0.81741  0.20981   2.1100  0.63073  0.20437  0.23977  3.5619E-002
0.58185
imember=   1  x2max=   3.1460   99.697   38.583   97.786   18.859
99.462   98.021   75.180   99.220   62.132   1.0765   1.7902   1.4551
2.3495
imember=   1  k2=   1.0332  0.11732  0.18000  0.82540  8.5563   2.2768
0.42388  0.20928   4.9932   1.4881   4.9867   4.9908  0.93315   4.9918
0.13158   1.2850   4.1625   4.9930  0.33685  0.37576   1.2531   4.9102
imember=   1  listprotatop=   7  19   4   22  11  23   9  23   4   22
12   4  17   4   1   18  14  13  16  18   8   23   2   4  20  17

imember=   1  prombelong=   1   2   3   1   4   2   4   5   6   3   2
7   6   8   4   5   9   4  10   3  11   8

imember=   2  k1=  1.0218E-002 -5.8965  1.0457E-002  1.0204E-002 -
0.37717  1.0378E-002 -0.18220  1.0418  1.0001E-002  1.6434E-002  1.4257
-2.5969  2.4823
imember=   2  k1'=  0.19810  0.69386  7.8144E-002   4.0744  0.64698
0.81741  0.20981   2.1100  0.63073  0.20437  0.23977  3.5619E-002
0.58185
imember=   2  x2max=   3.1460   99.697   38.583   97.786   18.859
99.462   98.021   75.180   99.220   62.132   1.0765   5.9486   1.4551
2.3495
imember=   2  k2=   1.0332  0.11732  0.18000  0.82540  8.5563   2.2768
0.42388  0.20928   4.9932   1.4881   4.9867   4.9908  0.93315   4.9918
0.13158   1.2850   4.1625   4.9930  0.33685  0.37576  0.57551   4.9102
imember=   2  listprotatop=   7  19   4   22  11  23   9  23   4   22
12   4  17   4   1   18  14  13  16  18   8   23  15  23  20  17

imember=   2  prombelong=   1   2   3   1   4   2   4   5   6   3   2
7   6   8   4   5   9   4  10   3  11   8
   costmin=   2.1780 costmean=   4.6843  coststddev=   11.469  end
generation=       11550


Solution 7

generation=       11750
```

```
imember=    1  k1= -5.0643E-002  7.9873 -0.85612  1.0016E-002  1.0063
0.59033  0.12060 -0.50596  1.0075E-002 -0.15827  1.0365  0.15984  1.0647
imember=    1  k1'=  4.8138E-002  0.14774  1.4561E-002   3.2654   1.1603
5.9282E-002  0.25782  0.42805  3.6209E-002  0.15972   2.1869   3.2170
3.2864
imember=    1  x2max=   99.810   6.0274   16.466   99.403   99.070
6.3695   91.877   98.717   18.021   23.662   99.566   98.127   96.613
2.6066
imember=    1  k2=   1.4330   3.1239   2.0248   4.2757   4.9547  0.70885
0.10839   1.7740  9.1358E-002  0.24743   4.9832   3.7686  0.77127
0.21932   1.1550  0.31922   1.9885  0.19694   2.9156   4.9770   1.0276
4.9604
imember=    1  listprotatop=  13  13   5    1  23  18  12  23   5    1
5   1  20   1   1    1  19  10  11  21  17   11   6  21   1   5

imember=    1  prombelong=   1   2   2    3   4   2   5   6   7    4   5
8   9  10   4    6  11   4  12   2   3   13

imember=    2  k1= -5.0643E-002  7.9873 -0.85612  1.0016E-002  1.0063
0.60363  0.12918 -0.50596  1.0075E-002 -0.15885  1.0023  0.15984  1.0152
imember=    2  k1'=  4.8138E-002  0.14832  1.4941E-002   3.2549   1.5610
6.1850E-002  0.28946  0.47054  3.6209E-002  0.15263   1.3499   3.2104
3.2864
imember=    2  x2max=   99.810   6.0274   16.466   99.403   99.070
6.3695   91.877   98.717   18.085   23.662   99.566   97.852   96.613
2.6066
imember=    2  k2=   1.4330   2.7688  0.40638  0.77896   4.9547  0.71579
0.32702   2.1661  0.92549  0.24293   4.9832   3.7887  0.77127  0.15968
1.5014  0.11337   1.9885  0.19694   2.9156   4.9770   1.0276   4.9604
imember=    2  listprotatop=  13  13   5    1  23  18  12  23   5    1
5   1  20   1   1    1  19  10  11  21  22   11   6  21   1   5

imember=    2  prombelong=   1   2   2    3   4   2   5   6   7    4   5
8   9  10   4    6  11   4  12   2   3   13
   costmin=   2.2933 costmean=   5.3868  coststddev=   8.9520  end
generation=       11750


Solution 8

generation=       11750
imember=    1  k1= -4.9288E-002  1.1012  1.6387E-002  1.0141E-002  2.3602
1.0041E-002  1.0001E-002  1.0273  6.9753E-002  5.0517E-002  1.0102
0.17294  1.0185
imember=    1  k1'=  2.5728E-002  0.30588  3.4854E-002  0.14387   3.0385
8.2261E-002  7.8199E-002  7.1625  1.7713E-002  0.31174   1.4611   3.0910
2.1502
imember=    1  x2max=   2.7805   9.2519  0.89543   9.9871   1.4184
9.9356   9.8973   9.8985   6.0377   3.7494   9.9690   6.9566   9.9355
9.4695
imember=    1  k2=   4.3701   1.9492  0.85586  0.36459  0.18230  0.11337
4.9825   1.6088   3.9267   1.0320   4.3911   4.5929   4.8668   2.1138
0.31902  0.48317  0.15363  0.38969  0.99009   3.3107  0.42667   4.4825
imember=    1  listprotatop=  23  17  11   16  18  23  12  23   2    1
11  15  12  23   9   15  22  13  12  23  22   13  11   7   4   7

imember=    1  prombelong=   1   2   2    3   4   2   4   5   6    2   7
8   6   9  10    5  11   4  12   2   1   13
```

```
imember=    1  prottypeact= F T T T T F T F T F T T F T F F F T T F T T
F

imember=    2  k1= -4.9288E-002  1.1012  1.6387E-002  1.0141E-002  2.3602
1.0041E-002  1.0001E-002  1.0273  6.9753E-002  5.0517E-002  1.0102
0.17294  1.0185
imember=    2  k1'=  2.5728E-002  0.30588  3.4854E-002  0.14387   3.0385
8.2261E-002  7.8199E-002  7.1625  1.7713E-002  0.31174   1.4611   3.0910
2.1502
imember=    2  x2max=   2.7805   9.2519  0.89543   9.9871   1.4184
9.9356   9.8973   9.8985   6.0377   3.7494   9.9690   6.9566   9.9355
9.4695
imember=    2  k2=   4.3701   1.9492  0.85586  0.36459  0.18230  0.11337
4.9825   1.6088   3.9267   1.0320   4.3911   4.5929   4.8668   2.1138
0.31902  0.48317  0.15363  0.38969  0.99009   3.3107  0.42667   4.4825
imember=    2  listprotatop=  23  17  11   16  18  23  12  23   2    1
11  15  12  23   9   15  22  13  12  23  22   13  11   7   4   7

imember=    2  prombelong=   1   2   2   3   4   2   4   5   6   2   7
8   6   9  10   5  11   4  12   2   1  13

imember=    2  prottypeact= F T T T T F F F T F T T F T F F
F T T F T T  F
   costmin=   2.2271 costmean=   6.2123  coststddev=   13.798  end
generation=       11750


Solution 9

generation=       11750
imember=    1  k1= -0.36623  4.8621 -6.1733E-002  1.0213E-002  4.0808
0.13586  1.0043E-002  2.6518  0.37715  1.0317E-002  0.31443 -0.10140 -
0.22402
imember=    1  k1'=  0.13593  0.29782  0.77513   1.0270  0.19600   2.8596
0.19553   2.2364   1.4979   1.7645   5.3782  0.83968  0.49686
imember=    1  x2max=   15.086   91.487   2.7825   70.944   2.5887
6.6934   98.851   78.557   75.489   99.478   50.877   99.427   98.366
2.6897
imember=    1  k2=  0.38780   2.8691   3.9070   4.1174  0.57170  0.85893
0.51724  0.36538   4.7496   2.1392   3.0897   4.9698   3.2737  0.10604
4.9651  0.59419  0.19340  0.30135  0.32289   4.9901   4.5288   4.7998
imember=    1  listprotatop=  19  23  12   23  19  23  12  23  12   23
20   4  20  18  12   19  17   2  22   9  13   23  18  23  15   9

imember=    1  prombelong=   1   2   2   3   4   5   4   6   7   2   2
7   8   9   4   6  10  11   9  12   1  13

imember=    1  prottypeact= F F T F T T T F F T T T T T F F
T F F T T T  F

imember=    2  k1= -0.36623  4.8621 -6.1733E-002  1.0213E-002  4.0808
0.13586  1.0043E-002  2.6518  0.37715  1.0317E-002  0.31443 -0.10140 -
0.22402
imember=    2  k1'=  0.13593  0.29782  0.77513   1.0270  0.19600   2.8596
0.19553   2.2364   1.4979   1.7645   5.3782  0.83968  0.49686
```

```
imember=   2  x2max=   15.086   91.487   2.7825   70.944   2.5887
6.6934   98.851   78.557   75.489   99.478   50.877   99.427   98.366
2.6897
imember=   2  k2=  0.38780   2.8691   3.9070   4.1174  0.57170  0.85893
0.51724  0.36538   4.7496   2.1392   3.0897   4.9698   3.2737  0.10604
4.9651  0.59419  0.19340  0.30135  0.32289   4.9901   4.5288   4.7998
imember=   2  listprotatop=  19  23  12   23  19  23  12  23  12   23
20   4  20  18  12   19  17   2  22   9  13   23  18  23  15   9

imember=   2  prombelong=   1   2   2   3   4   5   4   6   7   2   2
7   8   9   4   6  10  11   9  12   1  13

imember=   2  prottypeact= F F T F T T T F F F F T T F F T
T F F T F T  F
   costmin=   2.3500 costmean=   5.2291  coststddev=   8.8258  end
generation=       11750


Solution 10

generation=       11750
imember=   1  k1=  0.87354 -0.61584  1.0003E-002  0.20568
-1.2487  1.0204E-002  -2.0319 -3.2447  1.0067E-002  0.41088  1.0008 -
0.10185  1.0044
imember=   1  k1'=  2.5646E-002  8.9660  5.9346E-002   9.8369  0.38131
0.34151  5.8674E-002  4.8586   4.5348  0.12728   2.8454  0.75341
5.2644
imember=   1  x2max=   20.833   99.850   89.487   97.314   27.076
60.811   75.869   32.355   86.372   3.9077   99.422   90.475   5.2282
14.561
imember=   1  k2=  0.28185   4.9371  0.22187  0.27129  0.33001  0.36222
9.6169E-002   1.5883  0.27890   4.9872  0.66809   3.3153  0.84835
4.9964  0.34271   1.8395  0.81069  0.21127   4.9535   4.9524  0.38069
4.9744
imember=   1  listprotatop=  12  23  15   23   2  15   2  22   1   22
2  15  10  17  17   15  19  20   2  23  14   20  20  10  14  10

imember=   1  prombelong=   1   2   3   1   2   3   4   5   6   2   3
7   8   9  10   5  11   4  12   6   1  13

imember=   1  prottypeact= F T T F F T F T T F F T T T F F
F F T F F F  F

imember=   2  k1=  0.87354 -0.61584  1.0003E-002  0.20568
-1.2487  1.0204E-002  -2.0319 -3.2447  1.0067E-002  0.41088  1.0008 -
0.10185  1.0044
imember=   2  k1'=  2.5646E-002  8.9660  5.9346E-002   9.8369  0.38131
0.34151  5.8674E-002  4.8586   4.5348  0.12728   2.8454  0.75341
5.2644
imember=   2  x2max=   20.833   99.850   89.487   97.314   27.076
60.811   75.869   32.355   86.372   3.9077   99.422   90.475   5.2282
14.561
imember=   2  k2=  0.29699   4.9371  0.87777   2.4549  0.87525  0.22122
8.6933E-002  0.18331  0.87363   4.9872  0.42155   3.3153  0.84101
4.9964  0.34271   4.4286  0.81977  0.52467   4.9535   4.9524  9.5840E-
002   4.9744
imember=   2  listprotatop=  12  23  15   23   2  15   2  22   1   22
2  15  10  17  17   15  19  20   2  23  14   20  20  10  14  10
```

```
imember=    2  prombelong=    1    2    3    1    2    3    4    5    6    2    3
7    8    9   10     5   11    4   12    6    1   13

imember=    2  prottypeact= F T T F F T T T T F F T T T F T
F F T F F F  F
   costmin=    2.1147 costmean=    4.5859  coststddev=    16.858  end
generation=        11750


Solution 11

generation=        11750
imember=    1  k1=  1.8221E-002  2.4964  0.13111  1.0011E-002  0.34349
1.0016E-002  0.11820 -0.68587 -0.21202  1.0084E-002  5.1925  -1.1087 -
0.21515
imember=    1  k1'=  0.14823   9.9674   1.4305   8.2252  7.8194  0.12294
8.8063E-002  0.56706  8.8351E-002  0.76788  0.12596  0.93842  0.31017
imember=    1  x2max=  0.94770   87.815   34.000   99.933   1.2931
99.136   99.950   99.749   47.821   49.589   3.9650   99.789   99.212
0.48897
imember=    1  k2=  0.16014  7.8886  0.97200  0.88913  0.68481   4.0510
4.9987  0.13482   3.0576   4.9990  6.4240  0.42457   4.9949  0.73580
0.29662  0.17521  6.4196   1.4579  0.12596  0.71219  0.98504   4.9960
imember=    1  listprotatop=    5  12  14   22  15  22  10  22  13   14
10  15  23  15  23   14   9  18  23  22   6   16   7  13  14  18

imember=    1  prombelong=    1   2   3    1   4   2   4   5   6    7   2
8   6   9  10   11  12   4  13   3   1   12

imember=    2  k1=  1.8221E-002  2.4964  0.13111  1.0011E-002  0.34349
1.0016E-002  0.11820 -0.68587 -0.21202  1.0084E-002  5.1925  -1.1087 -
0.21515
imember=    2  k1'=  0.14823   9.9674   1.4305   8.2252  7.8194  0.12294
8.8063E-002  0.56706  8.8351E-002  0.76788  0.12596  0.93842  0.31017
imember=    2  x2max=  0.94770   86.963   35.780   99.933   1.2931
99.136   99.950   99.749   47.821   49.589   3.9650   99.789   99.212
0.48897
imember=    2  k2=   1.7956   1.4753  0.30043  0.22577  0.68481   4.0510
4.9987   1.9231   3.0576   4.9990  0.16076  0.42457   4.9949  0.73580
0.29662  0.17521   1.9480   1.4579  0.11513  6.0957  0.98128   4.9960
imember=    2  listprotatop=    5  12  14   22  15  22  10  22  13   14
10  15  23  15  23   14   9  18  23  22   6   16   7  13  14  18

imember=    2  prombelong=    1   2   3    1   4   2   4   5   6    7   2
8   6   9  10   11  12   4  13   3   1   12
   costmin=    2.2417 costmean=    4.3436  coststddev=    8.1630  end
generation=        11750
```

UCRL-JRNL-203451

Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The Special Case of Two Groups

J.R. Kercher , R.G. Langlois, B.A. Sokhansanj, C.F. Melius, J.N. Quong, F.P. Milanovich, B.W. Colston, Jr., K.W. Turteltaub, A.A. Quong

Lawrence Livermore National Laboratory, Livermore, California 94551

U.S. Department of Energy

Lawrence Livermore National Laboratory

This article was prepared for journal submission.

April 2004

DISCLAIMER

Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The Special Case of Two Groups

J. R.Kercher[1], R.G. Langlois[2], B.A. Sokhansanj[3],

C.F. Melius[4], J.N. Quong[4,5], F.P. Milanovich[6], B.W. Colston, Jr.[7],

K.W. Turteltaub[2], A.A. Quong[3,5]

[1]Environmental Sciences Division, L-235,

[2]Biodefense Division, [3]Chemical Biology and Nuclear Science Division,

[4]Chemistry and Chemical Engineering Division, [6]R Division, [7]M Division

Lawrence Livermore National Laboratory, P.O. Box 808

Livermore, California 94551 USA

[5]Georgetown University, Lombardi Cancer Center, 3970 Reservoir Road, NW,

Washington, DC 20057 USA

Corresponding author: J.R. Kercher, jkercher@llnl.gov, (925) 422-1416

Keywords: bio-signature, canonical variate analysis, clustered data, discrimination, Moore-Penrose, generalized inverse, biomarker, overfit

Running Head: Canonical analysis for singular within-group matrix

ABSTRACT

We suggest a new two-stage method for the canonical analysis (CA) of data classified into two groups (aka Fisher's linear discriminant analysis FLDA) for which the number of variables exceeds the number of observations. The first stage finds the minimal least squares solution to the canonical equation using the Moore-Penrose generalized inverse. We use the results of the first stage to rank and truncate the variable list for input to the second stage, a backward elimination (BE) CA. The three ranking criteria are based on sensitivities or correlations. We apply this technique to two trial examples: protein-expression in dialysis patients and gene-expression in tumors. We compare the results to a conventional two-stage canonical analysis consisting of a univariate signal-to-noise filter and a BE FLDA. In the first example, ferritin, interleukin 16, hepatitis-E virus (orf2_6KD), and α-fetoprotein comprise a *signature* or a set of variables that discriminates dialysis patients from controls ($P<<0.0001$). In the second example, the proposed technique successfully found combinations of variables (or signatures) that discriminate between tumor types ($P<<0.0001$) and had canonical variates with signal-to-nose ratios two to three times that of their average component. In both examples of class comparison, the statistics for performance (i.e., largest eigenvalue, Wilks ratio, average signal-to-noise ratio of signature variables) of the new two-stage, MPGI-based method exceed those for the conventional two-stage CA, indicating better discrimination and less tendency to overfit than the conventional FLDA. We find a nine-gene signature to distinguish two classes of lung primary tumors, defined by a 17-gene clustering of the tumors by Ramaswamy et al. (2003).

INTRODUCTION

Three types of problems and supervised methods. Three important problems in the multivariate analysis of gene- and protein-expression experiments are "class comparison", "class prediction", and "class discovery" (Simon 2003, Simon et al. 2003). Class discovery is the appropriate domain of the unsupervised methods of clustering (Simon 2003, Simon et al. 2003). Class comparison and class prediction are the domains of supervised methods (Radmacher et al. 2002). Dudoit et al (2002) suggest a similar tripartite division: "identification of new ... classes", "classification ... into known classes", and "identification of ‘marker’ [variables] that characterize the different … classes", which correspond to the class discovery, class prediction, and class comparison divisions, respectively, of Simon et al. (2003). Various supervised methods have been reviewed and compared by Dudoit et al. (2002). See Supplement Part A Section I for a partial list of supervised methods. Simon (2003), Simon et al. (2003), Radmacher et al. (2002), Ambroise and McLachlan (2002), and Nguyen and Rocke (2002) have discussed testing supervised methods. Some of the supervised methods are similar to each other or are either generalizations or special cases of other methods. For example, Simon (2003) notes that diagonal linear discrminant analysis (a Maximum Likelihood method, Dudoit et al. 2002) is similar to both the compound covariate predictor of Radmacher et al (2002) and weighted voting of Golub et al. (1999). Canonical analysis (CA) (Seal 1964), canonical variate analysis (Krzanowski 2000), and canonical discriminant analysis (Albert 1990) are synonyms for the multi-group generalization of Fisher's linear discriminant analysis (FLDA). Alternatively, FLDA is the special case of CA applied to two groups (e.g., Srivastava 2002 p. 258), as shown explicitly in Suppl.A Sec. I.

The overfit problem. A common thread that runs through discussions of class prediction and class comparison for expression data sets is that the number of variables p is typically much larger than the number of observations N (e.g., Radmacher et al, 2002, Dudoit et al. 2002). Simon et al. (2003) point out that for $p>>N$ "overfit" can occur, i.e., "the model fits random variations within the original data that do not represent true relationships that hold for independent data". Overfit is a form of "capitalization on chance" (e.g., Harris 2001). Some authors refer to $p>>N$ as a *supersaturated design* (e.g., Li and Lin 2002, Westfall et al. 1998).

Fisher's linear discriminant analysis and overfit. FLDA is often used to discriminate two groups (Xiong et al. 2000, Xiong et al. 2001a, Dudoit et al. 2002). Fisher (1936) originally devised FLDA as a non-probabilistic, geometrical, data-based analysis. Since then it has been shown that for normally distributed data, FLDA is the best predictor of sample membership for many criteria (Krzanowski 2000 p. 356, Jobson 1992, p. 258). However, Dudoit et al. (2002) note that the FLDA function is singular for $p>(N-2)$ and they find that FLDA is prone to overfit. They find that restricting the input variables to a small set improves class prediction of FLDA.

Current methods to address the overfit problem. A common approach to avoid overfit is to limit the number of variables used in the discrimination algorithm. Finding the subset to use is referred to as variable selection or subset selection. Stepwise techniques (i.e., stepwise, forward selection FS, backward elimination BE) have been used for variable selection for discrimination and regression (e.g., McCabe 1975, Hawkins 1976, Habbema and Hermans 1977, McHenry 1978, Westfall et al. 1998) and have been reviewed and compared (Hocking 1976, Costanza and Afifi 1979, McKay and

Campbell 1982, Krzanowski 2000). Both FS (Xiong 2001a, Xiong 2001b, Ambroise and McLachlan 2002, Quong et al. 2004), and BE algorithms (Ambroise and McLachlan 2002, Kozak et al. 2003) have been applied in expression studies. Unfortunately stepwise techniques for CA were originally designed for $p \le (N - h)$ where h is the number of groups, and when applied to supersaturated designs just by themselves, can still lead to overfit. We demonstrate an example of this in the Results section.

Because the stepwise procedure by itself produces overfit if applied to the original set of variables, often the variable list is first shortened using some univariate statistic to select only the most desirable variables for input to the discriminant analysis procedure. The selection statistics include the highest univariate signal-to-noise ratio (S_x, e.g., Ramaswamy et al. 2003), also known as prediction strength (Golub et al. 1999, Xiong et al. 2001a), the BW ratio of Dudoit et al. (2002), *t*-tests (e.g., Xiong et al. 2001a, Kozak et al. 2003), or Wilcoxin tests (e.g., Kozak et al. 2003). We refer to these methods as two-stage procedures in which stage one is a ranking procedure, based on a univariate statistic, followed by truncation. Stage two is a stepwise procedure on the truncated list.

A new method. In this paper, we suggest a new two-stage approach for CA for the special case of two groups (i.e., FLDA) and apply it to class comparison; in a companion paper, Kercher et al (2004) describe a generalization for the general case of any number of groups and apply the method to class prediction. We have restricted the current paper to the special case of two groups because FLDA is an important technique with many desirable properties, many practical applications have only two groups, and for two groups we can prove or derive special results such as analytic expressions for scaling properties, which give insight into the new method's performance.

The novel aspect of our approach is that we use the results of an approximate multivariate solution to rank variables in the first stage, rather than a univariate statistic as in conventional two-stage procedures. In the case of two groups, the first stage of the new method is a Moore-Penrose generalized inverse solution to the FLDA (MPGICA). The second stage is a stepwise FLDA, either backward elimination (BECA) or forward selection (FSCA).

We have two goals in this paper. First, we describe the MPGI solution to the CA problem for two groups and the two-stage procedure. Second we show the analyses of class comparisons for two example data sets (one protein- and one gene-based) in which we compare the results of the new method with the results of a conventional two-stage stepwise FLDA, i.e., BW-BECA. In the first stage in the conventional method, we use the BW ratio of Dudoit et al. (2002) to rank and truncate the variable list. The second stage is the stepwise FLDA. In the protein data set, we compare the new method with both the conventional two-stage stepwise FLDA (BW-BECA) and a conventional one-stage stepwise FLDA (FSCA-only), i.e., no pre-filtering with the BW ratio.

We note that Ambroise and McLachlan (2002) suggest that BE may incur less bias than FS. Except for one comparison to a pure FS exercise, we only present results for BE.

EXPERIMENTAL PROCEDURE AND DATA

Rules-Based Medicine™ (RBM) protein data of dialysis patients

Dialysis patients are frequently subject to inflammation, malnutrition, and cardiovascular problems (Sezer et al 2002, Don and Kaysen 2000, Zimmermann et al 2000, Freemont 2002). Various workers have observed that these conditions give rise to

several markers, which distinguish this population from the general population (Beerenhout et al 2002, Drueke and Massy 2002, Borawski et al 2002, Hung et al 2002, Level et al 2001, Massy and Nguyen-Khoa. 2002, Panichi et al 2001, Kato et al 2002, Schwedler et al 2001, Schwedler et al 2002, Stenvinkel et al 2000). Furthermore, the dialysis treatment itself seems to have biomarkers associated with it (Freemont 2002, Ikizler et al 2002, Tetta et al 2001, Memoli et al 2002, Schindler et al 2000, Horl 2002).

RBM used Luminex™ technology (reagent-coated fluorescent micro-spheres) to detect levels of 165 different proteins or protein-associated ligands in blood. Blood samples were obtained with informed consent from nineteen normal volunteers to constitute a control group and from eight dialysis patients to constitute a treatment group. The blood samples were processed to yield frozen aliquots of serum as described previously (Langlois et. al 2004). Frozen samples were coded before shipment to RBM. so that the experimentalists were blinded as to which samples were treatment vs. control. Our goal was to find a set of variables, which we refer to as a signature, which discriminates dialysis patients from control subjects.

Microarray gene expression data of primary and metastatic tumors.

An important problem in cancer research is the origin of the metastasis process. In particular, a fundamental question is whether signatures might exist that can predict the likelihood of metastasis of primary tumors. Ramaswamy et al. (2003) suggest that such signatures do exist and have proposed a 17-gene candidate signature. They analyzed an Affymetrix Hu6800/Hu35KsubA-microarray data set (Dataset A) of 64 primary and 12 metastatic tumors from six sites of origin (breast, lung, colorectal, prostate, uterus, and ovary) to find the 128 genes with the highest signal-to-noise ratios S_x

between the primary and metastatic tumors. Mapping these 128 genes to the corresponding 169 genes in an Affymetrix U95A-microarray data set (Dataset B) of lung primary tumors, they find a proposed 17-gene signature for discriminating between "primary-type" and "metastatic-type" primary tumors. We performed CA's on the 128-gene data set of the 64 primary and 12 metastatic tumors and the 169-gene data set of lung primary tumors.

THE CANONICAL EQUATION

If data has been classified into discrete groups (e.g., control, treatment 1, treatment 2, etc.), then canonical analysis is a multivariate technique designed to find the linear combination of the original variables that optimizes the separation of the groups.

The between-group variance-covariance matrix and the within-group variance-covariance matrix. The original data is represented by the matrix **X** with elements x_{ij}, which are the values of the variable i of observation or experimental unit j where $i \le p$, $j \le N$. Let $\mathbf{x}_s^j$ denote the jth column of **X**. The matrices **W** and **B** are the within-group and between-group sum-of-squares-and-cross-products SSCP matrices, respectively. See Suppl.A Sec. I for additional notation and definitions

The canonical equation. Based on geometric arguments, Seal (1964) derives the canonical equation (eq. 2) by finding a transformation to a new coordinate system (eq. 1), which maximizes the between-group variance while holding the within-group variance constant (eq. 3) along each axis in the new space. Krzanowski (2000, p.294, p.370) derives the canonical equation by introducing a transformation to the univariate case and then appealing to the union-intersection principle to maximize the F value given by the ratios of the between-group variance to the within-group variance. Mardia et al (1979, p.

338) derive the canonical equation by applying the likelihood ratio principle to the null hypothesis H_0: the $\boldsymbol{\mu} + \boldsymbol{\tau}_k$ lie in an *r*-dimensional hyperplane.

Let **E** be the matrix representing the transformation with matrix element E_{ij} and let $\mathbf{e}^i$ be the *i*th column vector of **E**. The canonical variable **y** in the new space is given by $\mathbf{y} = \mathbf{E}^{\mathbf{T}}\mathbf{x}$ where the superscript **T** designates the transpose or

$$y_{ij} = \sum_{k=1}^{p} x_{kj} E_{ki} \tag{1}$$

for the new canonical variate *i* and observation *j*. The canonical equation for $\mathbf{e}^i$ is

$$(\mathbf{B} - \lambda_i \mathbf{W})\mathbf{e}^i = 0 \tag{2}$$

There are *p* of these equations, one for each eigenvalue. The normalization equations are

$$\mathbf{e}^{i\mathbf{T}} \mathbf{W} \mathbf{e}^i / (N - h) = 1. \tag{3}$$

These equations fix the scale by normalizing the eigenvectors. In the standard canonical analysis, one solves eqs. 2 and 3 and then uses eq. 1 to find the values of the observations in the new coordinates. If the number of degrees of freedom of the within-group variance (*N*–*h*) exceeds *p*, then **W** is non-singular and eq. 2 can be solved by standard means. However, in both of the example data sets to be analyzed here, the number of variables exceeds the number of observations *p*>(*N*–*h*), **W** is singular, semi-positive definite, and $\mathbf{W}^{-1}$ does not exist. The singular nature of **W** rules out using standard techniques to solve eq. 2, and we analyze the canonical equation with a different approach.

GENERALIZED INVERSE CA FOR TWO GROUPS AND *p*>(*N*–*h*).

Moore-Penrose solution to the canonical equation for two groups.

Singular value decomposition of **W** *and* **B**. To analyze eq. 2 when *p*>(*N*–*h*), we first apply a singular value decomposition SVD on the matrices **W** and **B**. Because **B** and

W are symmetric, the SVDs are given by $\mathbf{V}^{\mathbf{T}}\mathbf{W}\mathbf{V} = \mathbf{\Delta}$ and $\mathbf{S}^{\mathbf{T}}\mathbf{B}\mathbf{S} = \mathbf{\Xi}$ where the columns of the $p \times p$ matrices **V** and **S** are denoted by $\mathbf{v}^i$ and $\mathbf{s}^i$, respectively. The matrices **V** and **S** are non-singular and orthogonal. The matrices **Δ** and **Ξ** are diagonal with the non-negative eigenvalues δ_i and ξ_i, respectively, in descending order on the diagonal. Typically, the number r of non-zero eigenvalues of **W** is given by $r=(N-h)$. However, $r<(N-h)$ is possible, but unlikely. The number of non-zero eigenvalues of matrix **B** is usually the number of degrees of freedom $(h-1)$ of **B**.

Generalized inverse approach. For the case of two groups, the matrix **B** has one non-zero eigenvalue, ξ_1 with eigenvector $\mathbf{s}^1$. The p-vector $\mathbf{e}^1$ can be expanded as a sum of the eigenvectors of **B**. That is, to find a vector **a** with components a_i such that $\mathbf{e}^1 = \mathbf{S}\mathbf{a}$, multiply both sides by $\mathbf{S}^{\mathbf{T}}$. Because **S** is orthogonal, we find $\mathbf{a} = \mathbf{S}^{\mathbf{T}}\mathbf{e}^1$. So **B** acting on $\mathbf{e}^1$ produces $\mathbf{B}\mathbf{e}^1 = \sum_i a_i \mathbf{B}\mathbf{s}^i = a_1\xi_1\mathbf{s}^1$. Thus, in the case of two groups, eq. 2 has the form $a_1\xi_1\mathbf{s}^1 = \lambda_1\mathbf{W}\mathbf{e}^1$ where a_1, λ_1, and $\mathbf{e}^1$ are unknowns. Because we will renormalize the solution for $\mathbf{e}^1$ by eq. 3, without loss of generality we can absorb a_1, ξ_1, and λ_1 into $\mathbf{e}^1$ and the canonical equation (eq. 2) becomes

$$\mathbf{W}\mathbf{e}' = \mathbf{s}^1 \qquad (4)$$

We will use the Moore-Penrose generalized inverse to solve eq. 4 because the MPGI is unique, exists for all matrices, and is well-studied with many useful properties. In particular, $\mathbf{A}^{+}\mathbf{b}$ is the minimal least squares solution to $\mathbf{A}\mathbf{x}=\mathbf{b}$ (e.g., Campbell and Meyer 1979 Thm 2.1.1.), where $\mathbf{A}^{+}$ is the MPGI of **A**. See Suppl.A Sec. II for this theorem and other properties of the Moore-Penrose inverse (Schott 1997). If eq. 4 is *consistent*, i.e.,

having an exact solution, then the MPGI also gives that solution. Define the $p \times p$ diagonal matrix **R** with elements R_{ij} as $R_{ii} = \delta_i^{-1}$ for $\delta_i > 0$, R_{ii}=0 for $\delta_i = 0$, and R_{ij}=0 for $i \neq j$. The matrix **R** is the MPGI of **Δ**. Because **W** is a symmetric matrix, the MPGI of **W** is given by $\mathbf{W}^+ = \mathbf{VRV}^{\mathbf{T}}$. The minimal least squares solution to eq. 4 is given by

$$\mathbf{e}' = \mathbf{W}^+\mathbf{s}^1 = \mathbf{VRV}^{\mathrm{T}}\mathbf{s}^1. \qquad (5)$$

The final form of vector $\mathbf{e}^1$ is found by renormalizing $\mathbf{e}'$ using eq. 3. Once $\mathbf{e}^1$ is found, eq. 2 yields $\lambda_1 = \mathbf{e}^{1\mathbf{T}}\mathbf{B}\mathbf{e}^1 / \mathbf{e}^{1\mathbf{T}}\mathbf{W}\mathbf{e}^1 = \mathbf{e}^{1\mathbf{T}}\mathbf{B}\mathbf{e}^1 / (N-h)$. So λ_1 specifies group separation.

Special properties of the generalized inverse canonical analysis.

A family of solutions. In the standard canonical analysis (nonsingular **W**) there is one unique solution. In the generalized inverse solution to the CA problem with singular **W**, there is a family of infinitely many solutions. In Suppl.A Sec. III, we show that all solutions to eq. 4 can be written in the form $\mathbf{e}' = \mathbf{VRV}^{\mathbf{T}}\mathbf{s}^1 + \mathbf{VYV}^{\mathbf{T}}\mathbf{s}^1$ where **Y** is a diagonal matrix whose *r* upper left diagonal entries Y_{ii} are all zero and whose other diagonal entries are arbitrary. We also show that the solution that we use in eq. 5 is the most conservative of all solutions for which $Y_{ii} \geq 0$.

Dependence of solutions on variability of units or changes in scales. Multivariate techniques are often sensitive to the relative scale of the individual variables. For example, Hoppner et al. (1999, p. 8) point out that cluster analysis is sensitive to the scales of the variables because of its reliance on distance measures. Also, principal components analysis, a multivariate technique for ungrouped data, is sensitive to the relative scales used for the individual variables (e.g. Morrison 1990 p. 314, Krzanowski 2000 p. 66, Mardia et al. 1979 p.219). The standard CA, which uses a nonsingular **W**

matrix, is different from these other multivariate techniques in that the results are independent of the relative scales of the variables (e.g. Mardia et al. 1979 p. 344). In Suppl.A Sec. IV, we demonstrate this result and that this argument does not extend to the MPGI solutions of the non-standard CA with singular **W**. That is, for the MPGI solution for the singular **W** case, the results depend on the units chosen for each variable. In Suppl.A Sec. V, we show that the eigenvalues in the MPGI solutions depend on the relative scale change.

Consider the following change in scale. We pick a specific variable *i* and rescale it (multiply all observed values of variable *i*) by a parameter (1+*k*); all other variables are left as is. Define $G_i(k)$ to be the fraction of the distance between the two groups due to variable *i* at scale factor (1+*k*). In Suppl.A Sec. VI, we give an analytical formula for the dependence $G_i(k)$ on *k* using theorems from Meyer (1973). While complicated, this formula for $G_i(k)$ indicates that the generalized inverse method can give preference to variables with larger scales. The dependence of $G_i(k)$ on *k* will depend on the covariance structure of the data (i.e., **W** and $\mathbf{W}^+$ matrices) and the vector between the group means. This formula indicates that $G_i(k)$ is the ratio of two fifth order polynomials in *k*. Thus, $G_i(k)$ may increase with *k* initially, but eventually $G_i(k)$ will saturate with *k* and asymptotically approach a limiting value. In practice with real data, we find $G_i(k)$ approaches an asymptotic value as *k* increases and that this value is usually positive and usually less than 1, but can be negative. We find empirically that the asymptotic limit for

the contribution of a variable is partially correlated with the univariate signal-to-noise ratio S_x for the variable.

In Suppl.A Sec. VII, we derive an expression $G_i(0)$ for the fraction of the distance between the group means that is due to variable *i* for no scale change as a function of the group means and $\mathbf{W}^+$. The denominator in the expression is the generalized Mahalanobis distance between the group means and the numerator is the fraction of generalized Mahalanobis distance due to variable *i*. If all variables have identical univariate within-group variance, then the formula for $G_i(0)$ is the product of S_x for *i* times a linear combination of all the signal-to-noise ratios. This linear combination is different for each variable, making it difficult to predict an overall relationship. We find in the Results that the univariate S_x accounts for about 43 percent of the variance of $G_i(0)$ in the dialysis data set. The univariate S_x is but one factor influencing $G_i(0)$; the multivariate correlations within the data are also important.

DESCRIPTION OF THE TWO-STAGE CA FOR $p>(N–h)$

New two-stage procedure: using MPGICA results in the stepwise algorithm

We use the generalized inverse canonical analysis results to rank the variables in the first stage of the two-stage procedure. Because the MPGI solution to eq. 2 is the best solution in the sense of least squares, we want to select those variables that are most important in the MPGI solutions for inclusion in the subsequent second stage BECA. After ranking, we truncate the list of variables to the number of available degrees of freedom ($N–h$) for use in the stepwise CA algorithm. We use three different criteria for ranking the variables. In the Results section we will compare the results of this two-stage

procedure for all three ranking criteria to the results of a conventional two-stage procedure. Algorithms to compute ranking criteria are given in Suppl.A Sec. VIII.

Criterion A: Sensitivity of the group separation to changes coefficients of $\mathbf{e}^1$. In this criterion we assume that those variables for which the Wilks ratio in canonical space is most sensitive are the most important. We use the sensitivity Γ_i of the Wilks ratio to the *i*th variable to rank the variables.

Criterion B: Absolute sensitivity. For this criterion, we sort on $|\Gamma_i|$, i. e., large changes, either increases or decreases, in the group separation may be important.

Criterion C: Correlation. For this criterion, we assume that the correlation of the original variable with the canonical axis produced by the MPGI CA determines its importance.

We designate the new two-stage procedure as MPGICA-X-BECA for backward elimination following the MPGI solution where X denotes criteria A, B, or C .

Conventional Two-Stage Procedure

The first stage of the conventional two-stage procedure consists of a ranking and truncation algorithm. We rank using the BW ratio of Dudoit et al. (2002). The second stage is a stepwise canonical analysis, here BECA. We designate the conventional two-stage procedure as BW-BECA (or equivalently BW-BEFLDA).

Inference Tests for the Stepwise Algorithms

Mathematical details of the inference tests can be found in Suppl.A Sec. IX.

Significance of eigenvectors (*canonical axes*). For each canonical analysis having *n* nonzero eigenvalues, we perform *n* tests on the significance of the eigenvalues and

eigenvectors (Cooley and Lohnes 1971 p. 249, Mardia et al. 1979 p. 343). We use this test as a first estimate as to which canonical axes we can discard as not significant.

Significance of individual variables: coefficients of eigenvectors. We test for significance of each coefficient of the significant eigenvectors in the stepwise canonical analyses. Thus, we test for the significance of either the added or the eliminated variable in the FS or BE procedure, respectively. Rao (1970) formulated this test for two groups. Hawkins (1976) and McHenry (1978) extended this test to an arbitrary number of groups. We denote the *P*-value for this test as $P_{Ft}(i)$ where *i* is the number of variables used in the CA. For brevity, refer to this test as RHM. McLachlan (1980) has analyzed the error rate for this test. Hawkins (1976) suggested the use of Bonferroni adjustments for this test. See Suppl.A Sec. IX for further discussion of the Bonferroni adjustment of $P_{Ft}(i)$.

Confidence intervals for the group means. We provide confidence intervals for the group means in canonical space (Mardia et al. 1979 p. 345, Seal 1964 p. 137).

Separation of the groups. Pairwise comparison. We conduct a pairwise comparison of the means of all the groups. This test is a Scheffe comparison *F* test on the group separation (e.g., Hays 1994, p. 455) and *post hoc* on the variables, i.e., canonical axes, to be used (Harris 2001, p. 222).

Overall group differences. We use Wilks ratio Λ as one factor in deciding between the stepwise CAs in the case of an equal number of variables for all three criteria. To compensate for an unequal number of variables, we use the rank of the significance (*P*-value) of Wilks Λ using the *F* approximation for its distribution due to Rao (1965, p. 471). Denote the *P*-value found from this test as $P_{FW}(l)$ where *l* is the number of variables used.

Implementation of the MPGI CA and the Stepwise Algorithms

Ideally, all variables should be measured in units in which the variance is identical across all variables. If there is no prior information on variance and all variables are measured in the same units, we default to the native units; otherwise we standardize all variables using the univariate within-group variance. We give implementation details in Suppl.A Sec. IX for the MPGI CA, the FSCA, and the BECA.

RESULTS AND ANALYSIS

Rules-Based Medicine, Inc.™ protein data of dialysis patients

Pretreatment of the data. The raw data consists of 161 protein measurements with a mix of units including both absolute values of concentration (e.g., mg/mL, μg/mL, ng/mL, etc) and relative scales using ratios with standards. We first standardize the data by setting the units for each variable such that the univariate within-group variance equals 1. See Supplement Part B Section X for details on pre-processing of the data.

Effect of scale on MPGI solutions. For each variable in turn, we solve the canonical equation using the MPGI for a variety of values of k to find the dependence of $G_i(k)$ on k where $(1+k)$ is the scale factor multiplying the variable for all observations. Recall that in the standard canonical analysis (i.e., nonsingular $\mathbf{W}$, $p \le (N-h)$) the factor k has no effect on $G_i(k)$. In Fig. 1, we show a plot of $G_i(k)$ as a function of k for the two variables, ferritin and stem cell factor, which have the largest signal-to-noise ratios S_x of 5.6 and 3.7, respectively, and have the two largest $G_i(k)$ at saturation. At low values of k, both variables contribute little to the group separation; at high values, both variables dominate the group separation. In Suppl.B Sec. X, we show plots of $G_i(k)$ for other

variables and tabulate the twenty variables with the highest maximum contributions to group separation, i.e., maximum $G_i(k)$ at high k. The correlation of the univariate S_x for each variable with max $G_i(k)$ and with $G_i(0)$ is 0.857 and 0.662, respectively, with an r^2 of 0.734 and 0.438, respectively.

Selecting canonical variates. We analyze the RBM dialysis data set with the MPGICA followed by backward elimination using all three criteria for ranking. For comparison, we also analyze the data with the conventional two-stage procedure, which is the BE algorithm pre-filtered by the BW ratio, BW-BECA. We show the $P_{Ft}(l)$ for rejecting the null hypothesis that the coefficient in $\mathbf{e}^1$ of the eliminated variable was significantly different from zero in Fig. 2a and Fig. 2b for MPGICA-A-BECA (Criterion A: sensitivity of Wilks ratio) and BW-BECA (conventional method), respectively. The last few variables eliminated are very significant. As we eliminate variables, eventually $P_{Ft}(l)$ drops below some pre-selected critical value α and stays below α. We set α=0.05. We refer to the first eliminated variable, after which the $P_{Ft}(l)$ stays below α, as the *cutoff* variable. In Fig. 2a and 2b, the cutoff variables are the fourth and the second variable, respectively. In Suppl.B Sec. XI, we show plots of the *P*-values $P_{Ft}(i)$ for the other criteria and tabulate the coefficients of $\mathbf{e}^i$ (*canonical signature*) and sensitivities of the Wilks ratio for the variables at cutoff for all MPGICA-X-BECA and BW-BECA. Note that all cases indicate that ferritin is the dominant variable.

Class comparison for new method and conventional method. In Table 1 we show the statistics for the results of the BE for the new two-stage method for all three criteria

and the conventional two-stage FLDA. The conventional method had the smallest eigenvalue λ_1, the highest Wilks Λ, and the highest ranking for the *P*-value of Wilks Λ at cutoff. Perhaps most importantly, the second variable chosen in the BW-BECA has the lowest S_x in Table 1, suggesting that the conventional CA may be the most prone to overfit. All of the statistics in Table 1 suggest that the conventional method produced the least desirable results. Finally, all values in at least one of the two groups were greater than the lower detection limit for all variables in Table 1, except for Parainfluenza_1. This too indicates a relatively poor performance by the conventional method.

We show boxplots for the 19-member control group and the eight-member group of dialysis patients for MPGICA-A-BECA and BW-BECA (conventional method) at cutoff in Fig. 3a and 3b, respectively. The gap between the medians of the groups for the new method is larger than that for the conventional two-stage BECA. In Suppl.B Sec. XI, we show boxplots for CA's for MPGICA-only, MPGICA-B-BECA and MPGICA-C-BECA. These plots show the improvement that the stepwise procedure adds to the MPGI solution. They also show that the new method, using either criterion B or C, produces superior group separation than the conventional method.

Analyzing the canonical variates. Even though the new two-stage BE procedure uses different variables at cutoff for the three criteria, the canonical axes produced by the new method for all three criteria are extremely similar. The axes from the new method are similar to the axis from the conventional method to a slightly lesser extent. In Suppl.B Sec. XI, we tabulate the correlations for the fifteen variables most positively correlated and the ten most negatively correlated with the canonical axis (*correlation signature*). These results indicate that the canonical axes generated by the new method

for all three criteria are almost identical and are similar to the axis from the conventional method to a slightly lesser extent. This suggestion is confirmed in Suppl.B Sec. XI where we tabulate the Spearman rank-correlations between each pair of lists of the 161 correlations of the variables with the canonical axis. For lists of 161 entries, these correlations are all highly significant ($P<<0.0001$).

Comparison to conventional pure forward selection (no BW pre-filter). We compare an MPGICA-A-BECA to a pure forward selection procedure with no pre-filter for the data of the dialysis study. See Suppl.B Sec. XII for procedure and results. The results in Table 2 suggest that more exploitation of random fluctuations (overfit) is occurring in the Bonferroni-adjusted FSCA than in the MPGICA-A-BECA.

Microarray gene expression data of primary and metastatic tumors.

Effect of scale. Ramaswamy et al. (2003) extracted from their Dataset A an 128-gene reduced data set with the best univariate S_x ratios, which separate primary from metastatic tumors. We use the MPGICA on the 128-gene data set, divided into primary tumors and metastases, to examine the effect of rescaling the expression for any one gene by the scale factor $(1+k)$. We plot the rescaled gene's contribution to the total distance between the two groups in Fig. 4. In Fig. 4a we show the fraction of the total distance on the canonical axis contributed by GenBank ID X82494, which reached the highest level of contribution of all 128 in the set. The maximum value for $G_i(k)$, 34%, is substantially lower than the 80% observed for ferritin in the dialysis data set. Recall that S_x for ferritin was 5.6 in the dialysis data set; by making its scale larger, the MPGI solution used mainly this one variable. In the tumor data set, the highest S_x was 0.707 for X82494. Hence, no

matter how large we make any one variable, several other variables contribute substantial distances to the total distance between the two groups at optimal separation. The variable with the second highest contribution to the total distance of separation is shown in Fig. 4b (GenBank ID S80437). There are several instances of negative contribution to the total distance. The maximal such case is GenBank ID X18900, which makes a negative contribution of about 19 percent of the total distance.

Class comparison: *Separation of primary from metastatic tumors in canonical space*. We performed the MPGICA-X-BECA and BW-BECA (conventional) analyses on the 128-gene data set divided into two groups: primary tumors and metastases. To identify CA's of interest, we find *P*-values $P_{Ft}(i)$, which meet a critical value of α=0.05, for the test on stepwise change in Wilks ratio at Bonferroni-adjusted cutoff, cutoff, and low-end values just above cutoff. We show comparative statistics of the CA's in Table 3. Table 3 indicates that the conventional method produces results that are less acceptable than the results from the MPGI method. Note also that the average S_x is slightly less for the conventional method than for the MPGI solutions, indicating that the conventional method may be more prone to overfit. We show boxplots for MPGICA-B-BECA and BW-BECA (conventional) for the Bonferroni-adjusted cutoffs in Fig. 5a and 5b, respectively, and for low-end variables in Fig. 5c and 5d, respectively. Adding genes to the set used for CA increases the separation of the groups relative to the size of the groups. For both the Bonferroni-adjusted and unadjusted cutoff, the medians for the MPGI solutions are farther apart than the medians for the conventional solutions. Good separation occurs for low-end variables. In Suppl.B Sec. XIII, we tabulate the $\mathbf{e}^1$ vector

transformations (canonical signatures) for the CA's, tabulate the 15 genes and ten genes with the highest positive and highest negative correlation, respectively, with the canonical axis, and show additional boxplots and graphs of the *P*-value $P_{Ft}(i)$. The set of genes with the highest absolute correlation (correlation signature) is a convenient tool to describe the gradient that the canonical axis represents.

Lung data. Signatures for groups classified by clustering. Ramaswamy et al. (2003) used the 128 genes, identified in Dataset A, to extract the 169 corresponding genes from Dataset B, the primary-tumor lung data for 62 patients, 31 with recurring and 31 with non-recurring tumors. Using clustering, Ramaswamy et al. found a collection of 17 genes, which divided the 62 patients into two groups of 38 patients (group 0) and 24 patients (group 1) that had significant overlap with the non-recurrent and recurrent groups, respectively. We used the conventional BW-BECA and the three MPGICA-X-BECA on this 169-gene data set grouped into cluster group 0 (38 patients) and the cluster group 1 (24 patients). The two smallest numbers of genes found that separate the two groups *with no overlap* were the nine genes (P_{Ft}=0.05) and eleven genes (P_{Ft}=0.15) from MPGICA-A-BECA and BW-BECA, respectively. See boxplots in Fig. 6. The results for the conventional method are inferior to those of the MPGI approach. The lists of genes, lists of coefficients of $\mathbf{e}^1$, and plots of $P_{Ft}(i)$ are given in the Suppl.B Sec. XIV.

DISCUSSION

Use of canonical analysis. In the standard canonical analysis, in which $p \leq (N-h)$ and **W** is nonsingular, one usually determines the significance of the axes, significance of the group separation, etc. In addition to these issues, in the case of $p > (N-h)$ and **W**

singular, because the number of variables in the list we draw from is so large, selecting from it without causing random fluctuations to generate plausible, yet incorrect, results is a serious problem. Consider how the proposed new method addresses overfit.

The nature of the MPGI-BECA. The generalized inverse gives the best overall approximate solution in the sense of least-squares to a set of canonical equations, in which the solution is forced to include all variables. We systematically order the variables using the MPGI results so that only the most relevant subset, defined by criteria A, B, or C, is used in the subsequent stepwise canonical analysis. That is, the MPGICA acts as a filter predicated on the property that the MPGI is the minimal least squares solution to the canonical equation. Or, alternatively, we can regard the BECA as a refinement of the MPGICA solution. Under the assumption that the variance is constant across all the variables, variables with the largest values will tend to dominate the results. This tendency is mitigated by correlations with other variables to produce the best overall solution in the sense of least squares. The variables that are discarded before the BE canonical analysis are discarded on their lack of contribution to the MPGICA results. In contrast the conventional method discards variables on a univariate basis, which does not take correlations with other variables into account.

We find that the new two-stage method using either criteria A, B, or C produces results superior to the conventional method for the Wilks ratio statistic, group separation, signal-to-noise ratio for the variables in the canonical signature, or rank of P_{FW}.

Interpretation of the canonical axes. Usually the single most important description of the canonical axes are those variables that are highly correlated with the axes. Often the canonical axes represent gradients along which one or more processes

change in importance. This is the basis on which we can assign meaning to canonical axes. By noting which variables change along these gradients, sometimes we can infer what underlying process (or processes) is associated with the axis. It is interesting to note that sometimes the variable that is most important (most highly correlated) on any one axis is not necessarily the variable used to construct the axis. Consider that the gene that is most correlated with the canonical axis in Suppl.B Sec. XIII is J03464. This gene was not used to construct the axes. Two types of *signatures* are those variables actually used in the canonical transformations (*canonical signatures*) or those variables highly correlated with canonical axes (*correlation signatures*). The former is discrete and more specific; the latter is robust with respect to ranking criteria.

Limitations of canonical analysis. Like all supervised methods, canonical analysis relies on external information of group classification to construct the discrimination function. While canonical analysis can aid in class discovery as an auxiliary technique, it does not replace clustering. Furthermore, CA results can be sensitive to outliers and CA does not always perform well if normality does not apply or if the covariance matrices are very different. However, despite these limitations, we have found it to be a useful tool that complements existing techniques.

Analysis of examples. The dialysis patients were well discriminated from the control subjects with relatively few protein-concentration variables. For this data set there are several biomarkers, i.e., single variables each of which individually distinguishes the two groups. The canonical analysis found signatures or combinations of such biomarkers that enhanced group separation. Different ranking criteria for variable selection for the stepwise procedures led to similar canonical axes. Unlike the dialysis

data, for which several proteins had S_x>1, all variables in the 128-gene tumor data set had signal-to-noise ratios less than 0.71. The primary tumors, taken as a single group, were well discriminated from the metastases; however, the tumor discrimination required many more variables than the dialysis data. Results for CA can differ for expression data depending on either the type of variables, the type of classification to which the data has been subjected, or the internal structure of the data. Combinations of genes (or signatures) in the tumor data had markedly better discrimination than individual genes. Canonical analysis discriminated the two types of primary tumors in the proposed classification of Ramaswamy et al. (2003). We found a nine-gene signature for the two types that was more parsimonious than the 17-gene signature proposed by Ramaswamy et al. (2003).

We find that both the statistics for discrimination and the signal-to-noise ratios of the canonical signatures for the new two-stage method are superior to those of the conventional method in the examples shown here. These results indicate that the new method is more useful for class comparison than the conventional method and shows promise to be less prone to overfit.

ACKNOWLEDGMENTS

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48. This work was supported by the LLNL Laboratory Directed Research and Development program.

SUPPLEMENT

Kercher, J.R. , R.G. Langlois, B.A. Sokhansanj, C.F. Melius, J.N. Quong, F.P. Milanovich, B.W. Colston, Jr., K.W. Turteltaub, A.A. Quong. 2004. Supplement to variable selection in canonical analysis of gene- and protein-expression data: the special case of two groups. Part A. Methods of analysis. http:/[TO_BE_DETERMINED]. Part B. Results. http:/[TO_BE_DETERMINED].

REFERENCES

Albert, J.M. 1990. Reduced-rank regression models for discriminant analysis of longitudinal data. Ph.D. Dissertation, University of Michigan. UMI.

Ambroise, C., G.J. McLachlan. 2002. Selection bias in gene expression on the basis of microarray gene-expression data. Proc. National Acad. Sciences 99:6562-6566.

Beerenhout, C., O. Bekers, J.P. Kooman, F.M. van der Sande, K.M.L. Leunissen. 2002. A comparison between the soluble transferrin receptor, transferring saturation and serum ferritin as markers of iron state in hemodialysis patients. Nephron 92: 32-35.

Borawski, J., K. Pawlak, M. Mysliwiec. 2002. Inflammatory markers and platelet aggregation tests as predictors of hemoglobin and endogenous erythropoietin levels in hemodialysis patients. Nephron 91:671-681.

Campbell, S.L., C.D. Meyer, Jr. 1979. Generalized inverses of linear transformations. Pitman: London.

Cooley, W.W., P.R. Lohnes. 1971. Multivariate data analysis. John Wiley & Sons.

Costanza, M.C., A.A. Afifi. 1979. Comparison of stopping rules in forward stepwise discriminant analysis. Journal of the American Statistical Association 74:777-785.

Don, B.R., G.A. Kaysen. 2000. Assessment of inflammation and nutrition in patients with end-stage renal disease. J. Nephrology 13:249-259.

Drueke, T.B., Z.A. Massy. 2002. Advanced oxidation protein products, parathyroid hormone and vascular calcification in uremia. Blood Purification 20: 494-497.

Dudoit, S., J. Fridlyand, T.P. Speed. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Statistic. Assoc. 97:77-87

Freemont, A.J. 2002. The pathology of dialysis. Seminars in Dialysis 15:227-231.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7:179-186.

Golub, R.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeck, J.P. Mesirov, H. coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286:531-537.

Habbema, J.P.E., J. Hermans. 1977. Selection of variables in discriminant analysis by *F*-statistic and error rate. Technometrics 19:487-493.

Harris, R.J. 2001. A primer of multivariate statistics. 3rd ed. Lawrence Erlbaum Assoc.:Mahwah, NJ.

Hawkins, D.M. 1976. The subset problem in multivariate analysis of variance. J. Royal Statist. Soc. B 38:132-139.

Hays, W.L. 1994. Statistics. 5th ed. Wadsworth Thomson Learning.

Hocking, R.R. 1976. The analysis and selection of variables in linear regression. Biometrics 32:1-49.

Hoppner, F., R. Kruse, F. Klawonn, T. Runkler. 1999. Fuzzy cluster analysis. Wiley.

Horl, W.H. 2002. Hemodialysis membranes: Interleukins, biocompatibility, and middle molecules. J. Am. Soc. Nephrology 13: S62-S71 Suppl. 1

Hung A.M. , G.M. Chertow , B.S. Young, S. Carey, K.L. Johansen. 2002. Inflammatory markers are unrelated to physical activity, performance, and functioning in hemodialysis. J. Renal Nutrition 12:170-176.

Ikizler, T.A., L.B. Pupim, J.R. Brouillette, D.K. Levenhagen, K. Farmer, R.M. Hakim, P.J. Flakoll. 2002. Hemodialysis stimulates muscle and whole body protein loss and alters substrate oxidation. AM. J. Physiology-Endocrin. & Metabol. 282: E107-E116.

Ip, W.C., H. Wong, Y. Li, Z. Xie. 1999. Threshold variable selection by wavelets in open-loop threshold autoregressive models. Stat. & Prob. Letters 42:375-392.

Jobson, J.D. 1992. Applied Multivariate Data Analysis: Vol. II. Categorical and multivariate methods. Springer-Verlag.

Kato, A, M. Odamaki, T. Takita, Y. Maruyama, H. Kumagai, A. Hishida. 2002. Association between interleukin-6 and carotid atherosclerosis in hemodialysis patients. Kidney International 61:1143-1152.

Kercher, J.R., J.N. Quong, K.J. Wu, A.A. Quong. 2004. Variable selection in canonical analysis of gene- and protein-expression data: the general case for multiple groups. (manuscript in preparation).

Kozak, K.R., M.W. Amineus, S.M. Pusey, F. Su, M.N. Luong, S.A. Luong, S.T. Reddy, R. Farias-Eisner. 2003. Identification of biomarkers for ovarian cancer using storn anion-exchange ProteinChips: potential use in diagnosis and prognosis. Proc. National Acad. Science 100:12343-12348.

Krzanowski, W.J. 2000. Principles of multivariate analysis. Clarendon Press, Oxford.

Langlois, R.G, J.E. Trebes, E.A. Dalmasso, Y. Ying, R.W. Davies, M.P. Curzi, B.W. Colston Jr., K.W. Turteltaub, J. Perkins, B.A. Chromy, M.W. Choi, G.A. Murphy, J.P. Fitch, and S.L. McCutchen-Maloney. 2004. Serum protein profile alterations in hemodialysis patients. Am. J. Nephrology (in Press)

Level, C., P. Chauveau, Y. Delmas, C. Lasseur, G. Pelle, E. Peuchant, D. Montaudon, C. Combe. 2001. Procalcitonin: a new marker of inflammation in haemodialysis patients? Nephrology Dialysis Transplantation 16:980-986.

Li, R., D.K.J. Lin. 2002. Data analysis in supersaturated designs. Stat. & Prob. Letters 59:135-144.

Mardia, K.V., J.T. Kent, J.M. Bibby. 1979. Multivariate analysis. Academic Press.

Massy, Z.A., T. Nguyen-Khoa. 2002. Oxidative stress and chronic renal failure: Markers and management. J. Nephrology 15: 336-341.

McCabe, G.P. 1975. Compuations for variable selection in discriminant analysis. Technometrics 17:103-109.

McHenry, C.E. 1978. Computation of a best subset in multivariate analysis. Appl. Statist. 27:291-296.

McKay, R.J. and Campbell, N.A. 1982. Variable selection techniques in discriminant analysis. I. Description. Br. J. Math. Statist. Psychol. 35:1-29.

Memoli, B., R. Minutolo, V. Bisesti, L. Postiglione, A. Conti, L. Marzano, A. Capuano, M. Andreucci, M.M. Balletta, B. Guida, C. Tetta. 2002. Changes of serum albumin and C-reactive protein are related to changes of interleukin-6 release by peripheral blood mononuclear cells in hemodialysis patients treated with different membranes. Am. J. Kidney Diseases 39: 266-273.

McLachlan, G.J. 1980. On the relationship between the F test and the overall error rate for variable selection in the two-group discriminant analysis. Biometrics 36:501-510.

Meyer, C.D. Jr. 1973. Generalized inversion of modified matrices. SIAM J. Appl. Math. 24:315-323.

Morrison, D.F. 1990. Multivariate statistical methods. 3rd ed. McGraw-Hill.

Nguyen, D.V., D.M. Rocke. 2002. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics 18:1216-1226.

Panichi, V., M. Migliori, S. De Pietro, D. Taccola, A.M. Bianchi, M. Norpoth, M.R. Metelli, L. Giovannini, C. Tetta, R. Palla. 2001. C reactive protein in patients with chronic renal diseases. Renal Failure 23:551-562.

Quong, J.N., K.J. Wu, J.R. Kercher, M. Knize, K. Kulp, A.A. Quong. 2004. A signature-based method to distinguish time-of-flight secondary ion mass spectra from biological samples. (Manuscript to be submitted).

Radmacher, M.D., L.M. McShane, R. Simon. 2002. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 9:505-511

Ramaswamy, S., K.N. Ross, E.S. Lander, T.R. Golub. 2003. A molecular signature of metastasis in primary solid tumors. Nature Genetics 33:49-54.

Rao, C.R. 1965. Linear statistical inference and its applications. John Wiley, New York.

Rao, C.R. 1970. Inference on discriminant function coefficients. P. 587-602. *In* Essays on Probability and Statistics (R.C. Bose, I.M. Chakravarti, P.C. Mahalanobis, C.R. Rao, K.J.C. Smith, eds) University of North Carolina Press:Chapel Hill, NC.

Schindler, R., O. Boenisch, C. Fischer, U. Frei. 2000. Effect of the hemodialysis membrane on the inflammatory reaction in vivo. Clinical Nephrol. 53 (6): 452-459.

Schott, J.R. 1997. Matrix analysis for statistics. John Wiley & Sons : New York.

Schwedler, S.B., T. Metzger, R. Schinzel, C. Wanner. 2002. Advanced glycation end products and mortality in hemodialysis patients. Kidney International 62: 301-310.

Schwedler, S., R. Schinzel, P. Vaith, C. Wanner. 2001. Inflammation and advanced glycation end products in uremia: Simple coexistence, potentiation or causal relationship? Kidney International 59:S32-S36 Suppl. 78.

Seal, H. 1964. Multivariate statistical analysis for biologists. Wiley.

Sezer S, Ozdemir FN, Arat Z, Turan M, Haberal M. 2002.Triad of malnutrition, inflammation, and atherosclerosis in hemodialysis patients. Nephron 91:456-462.

Simon, R. 2003. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. British Journal of Cancer 89:1599-1604.

Simon, R., M.D. Radmacher, K. Dobbin, L.M. McShane. 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classfication. Journal of the National Cancer Institute 95:14-18.

Srivastava, M.S. 2002. Methods of multivariate statistics. Wiley.

Stenvinkel, P. B. Lindholm, M. Heimburger, O. Heimburger. 2000. Elevated serum levels of soluble adhesion molecules predict death in pre-dialysis patients: association,vith malnutrition, inflammation, and cardiovascular disease. Nephrology Dialysis Transplantation 15:1624-1630.

Tetta, C., S. David, F. Mariano, C. De Nitti, V. Panichi. 2001. Alterations of the cytokine network in hemodialysis. J. Nephrology 14:S22-S29 Suppl. 4

Westfall, P.H., S.S. Young, D.K.J. Lin. 1998. Forward selection error control in analysis of supersaturated designs. Statist. Sinica 8:101-117.

Xiong M.M., L. Jin, W.J. Li, E. Boerwinkle. 2000. Computational methods for gene expression-based tumor classification. Biotechniques 29: 1264-1270.

Xiong, M. W. Li, J. Zhao, L. Jin, E. Boerwinkle. 2001a. Featgure (gene) selection in gene expression-based tumor classification. Molecular Genetics Metabolism 73:239-247.

Xiong, M., X. Fang, J. Zhao. 2001b. Biomarker identification by feature wrappers. Genome Research 11:1878-1887.

Zimmermann, J., T. Metzger, L. Schramm, C. Wanner. 2000. Causes and consequences of the chronic inflammatory state in chronic dialysis patients. Nieren-und Hochdruckkrankheiten 29:427-431.

Table 1. Statistics for results of the MPGI-based, two-stage canonical analyses compared to the results for a conventional two-stage stepwise CA with a univariate filter. Canonical analyses are for dialysis data for eight dialysis patients and 19 control volunteers. The fourth column is the rank of the *P*-value of the Wilks ratio had the variables been selected *a priori*.

Procedure/ Criterion	Canonical eigenvalue λ_1	Wilks ratio Λ	Rank of *P*-value of Wilks	Cutoff variables	Signal-to-noise ratio
MPGI CA-A-BECA	158.6	0.0063	2	Ferritin	5.59
				IL-16	2.27
				Hepatitis.E.Virus.orf2.6KD	1.55
				α-Fetoprotein	1.64
MPGI CA-B-BECA	104.4	0.0095	3	Ferritin	5.59
				Hepatitis.E.Virus.orf2.6KD	1.55
				Stem_Cell_Factor	3.74
MPGI CA-C-BECA	85.1	0.0116	1	Ferritin	5.59
				Hepatitis.E.Virus.orf2.6KD	1.55
BW-BE CA (conventional)	46.9	0.0209	4	Ferritin	5.59
				Parainfluenza_1	1.15

Table 2. Comparisons of correlations with the canonical axis and signal-to-noise ratios for variables selected by MPGI canonical analysis pre-filter for backward elimination CA and for those variables selected by Bonferroni-adjusted pure forward selection canonical analysis. Results for dialysis study using RBM protein data.

Variable from MPGICA-A-BECA	Correl-ation	Univariate Signal-to-Noise ratio	Variable from Bonferroni FSCA (conventional)	Correl-ation	Univariate Signal-to-Noise ratio
Ferritin	0.986	5.59	Ferritin	0.983	5.59
IL-16	0.926	2.27	Fibrinogen	0.432	0.525
Hepatitis_E_Virus(orf2_6KD)	0.855	1.55	Hepatitis_C_Core	0.504	0.568
α-Fetoprotein_	0.852	1.64	PCNA	0.301	0.309

Table 3. Comparison of statistics for MPGI-based, two-stage and conventional two-stage canonical analyses of gene-expression data for 128-gene experiment for primary tumors and metatstases (Ramaswamy et al. 2003). Column 7 is the signal-to-noise ratio of the resulting canonical variate. The maximum univariate S_x is 0.707 for this data set.

Procedure/ Criterion	λ_1	Wilks ratio Λ	Rank of $P_{FW}(l)$	Number of variables	Average univariate Signal-to-noise ratio	CA Signal-to-noise ratio
			Bonferroni-adjusted cutoff			
MPGI CA-A-BE CA	1.19	0.456	8	3	0.536	1.10
MPGI CA-B-BE CA	1.73	0.366	3	3	0.592	1.41
MPGI CA-C-BE CA	1.52	0.397	5	2	0.641	1.30
BW-BE CA (conventional)	1.00	0.500	9	3	0.511	1.00
			Cutoff			
MPGI CA-A-BE CA	1.58	0.387	5	5	0.531	1.30
MPGI CA-B-BE CA	1.73	0.366	3	3	0.592	1.41
MPGI CA-C-BE CA	1.81	0.356	2	3	0.591	1.39
BW-BE CA (conventional)	1.34	0.428	7	4	0.517	1.21
			Low-end			
MPGI CA-A-BE CA	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
MPGI CA-B-BE CA	3.27	0.234	1	11	0.563	1.93
MPGI CA-C-BE CA	2.49	0.286	4	8	0.572	1.60
BW-BE CA (conventional)	2.53	0.283	6	12	0.526	1.64

FIGURE CAPTIONS

Fig. 1. Plots of $G_i(k)$ as a function of k for dialysis data, where G_i is the fraction of the distance between the two groups in canonical space accounted for by variable i. The factor k multiplies the variable i in the **X** data matrix. (a) Ferritin is i=17. (b) Stem cell factor is i=53.

Fig.2. Plot of P-values $P_{Ft}(i)$ for dialysis data in rejecting the null hypothesis that the coefficient of the variable in the eigenvector of the canonical equation is not significantly different from zero. This test is from Rao (1970), Hawkins (1976), and McHenry (1978) for variables added or eliminated from canonical analyses. (a) MPGICA-A-BECA (backward elimination input pre-filtered by criterion A: [sensitivity of Wilks ratio] using results of MPGI solution). (b) BW-BECA (conventional backward elimination pre-filtered by univariate BW ratio).

Fig. 3. Boxplots of the two groups (eight dialysis patients, 19 control subjects) of RBM blood-serum protein data on canonical axis 1. Results are for (a) MPGICA-A-BECA (Moore-Penrose CA followed by ranking using criterion A:[Sensitivity of Wilks ratio] and truncation ofinput list for backward elimination CA) using cutoff CA at four proteins; (b) BW-BECA (conventional two-stage backward elimination pre-filtered by BW ratio), two proteins. The bottom of each box corresponds to the 25th percentile; the top of each box is the 75th percentile. The horizontal line within the box is the median. Any data point being a distance away from either the top or bottom by more than 1.5 times the distance between the top and bottom of the box is an outlier, designated by an open circle. The range of the remaining data is indicated by the horizontal bars.

Fig. 4. Fraction $G_i(k)$ of distance between the two groups of primary tumors and metastases in canonical space due to one variable as a function of the scale factor k multiplying that variable. The distances are those produced by the generalized inverse solutions to the canonical equation. In Figs. 4a and 4b, genes 117 and 89, respectively, are Affymetrix Hu6800/Hu35KsubA probes X82494_at, and S80437_s_at, respectively, corresponding to GenBank ID's X82494 and S80437, respectively.

Fig. 5. Boxplots of metastases and primary tumors following backward elimination. (a) Results for MPGICA-B-BECA at three-gene unadjusted (and Bonferroni-adjusted) cutoff step. (b) Results for BW-BECA at the three-gene Bonferroni-adjusted cutoff step. (c) Results for MPGICA-B-BECA at the eleven-gene low-end step. (d) Results for BW-BECA at the low-end twelve-gene step. Data from Ramaswamy et al. (2003) for 128 genes from their Dataset A for 64 primary tumors and 12 metastases.

Fig. 6. Result of backspace elimination canonical analysis at nine genes and eleven genes in (a) and (b), respectively for lung primary data from 169-genes from Dataset B. The groups were those designated as cluster 0 or cluster 1 by Ramaswamy et al. (2003). The groups are the result of a clustering procedure. (a) MPGICA-A-BECA. (b) BW-BECA (conventional).

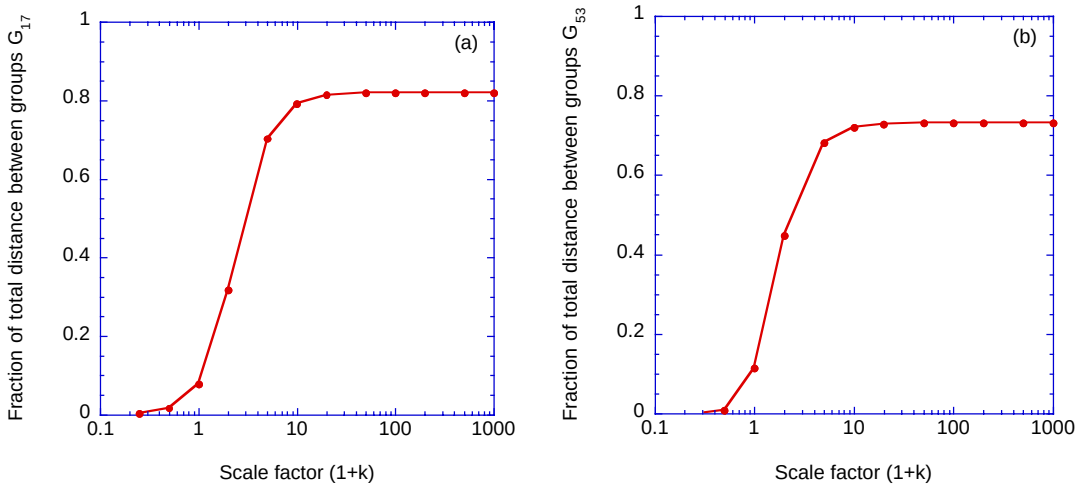

Fraction of total distance between groups G_{17}
(a)
Scale factor (1+k)
Fraction of total distance between groups G_{53}
(b)
Scale factor (1+k)

Fig. 1.

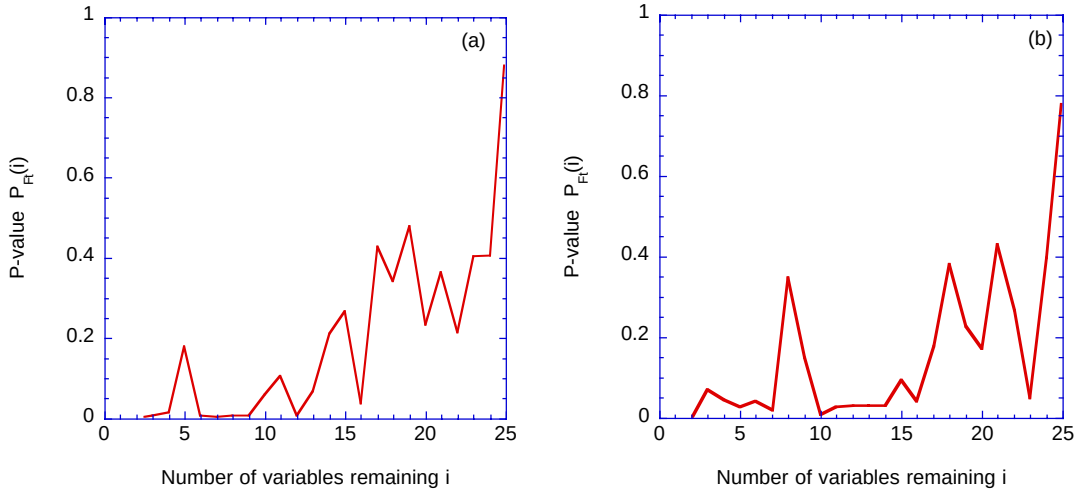

(a)
(b)
P-value $P_{FI}(i)$
Number of variables remaining i
0
0.2
0.4
0.6
0.8
1
0
5
10
15
20
25

Fig. 2

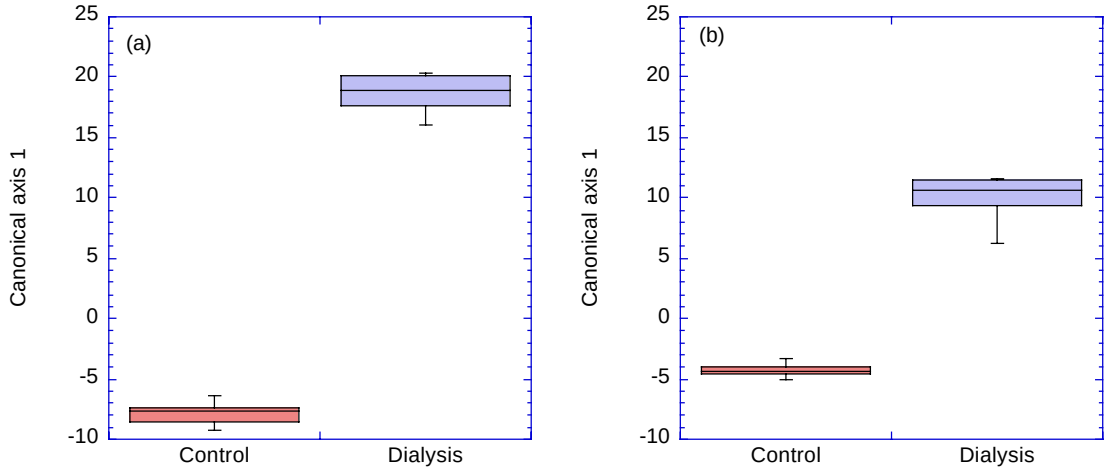

(a)
(b)
Canonical axis 1
Canonical axis 1
25
20
15
10
5
0
-5
-10
Control
Dialysis
Control
Dialysis

Fig. 3

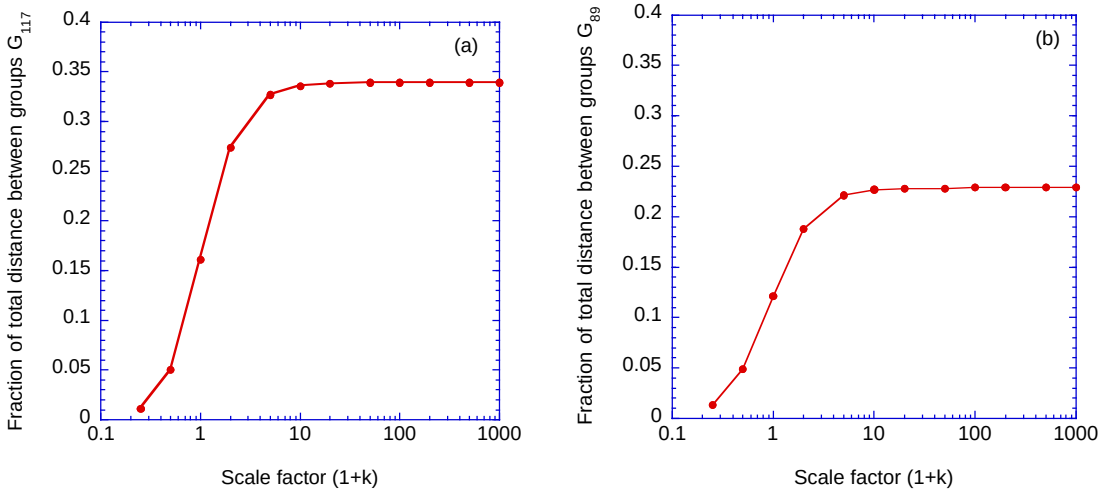

Fraction of total distance between groups G_{117}
(a)
Scale factor (1+k)
Fraction of total distance between groups G_{89}
(b)
Scale factor (1+k)

Fig. 4

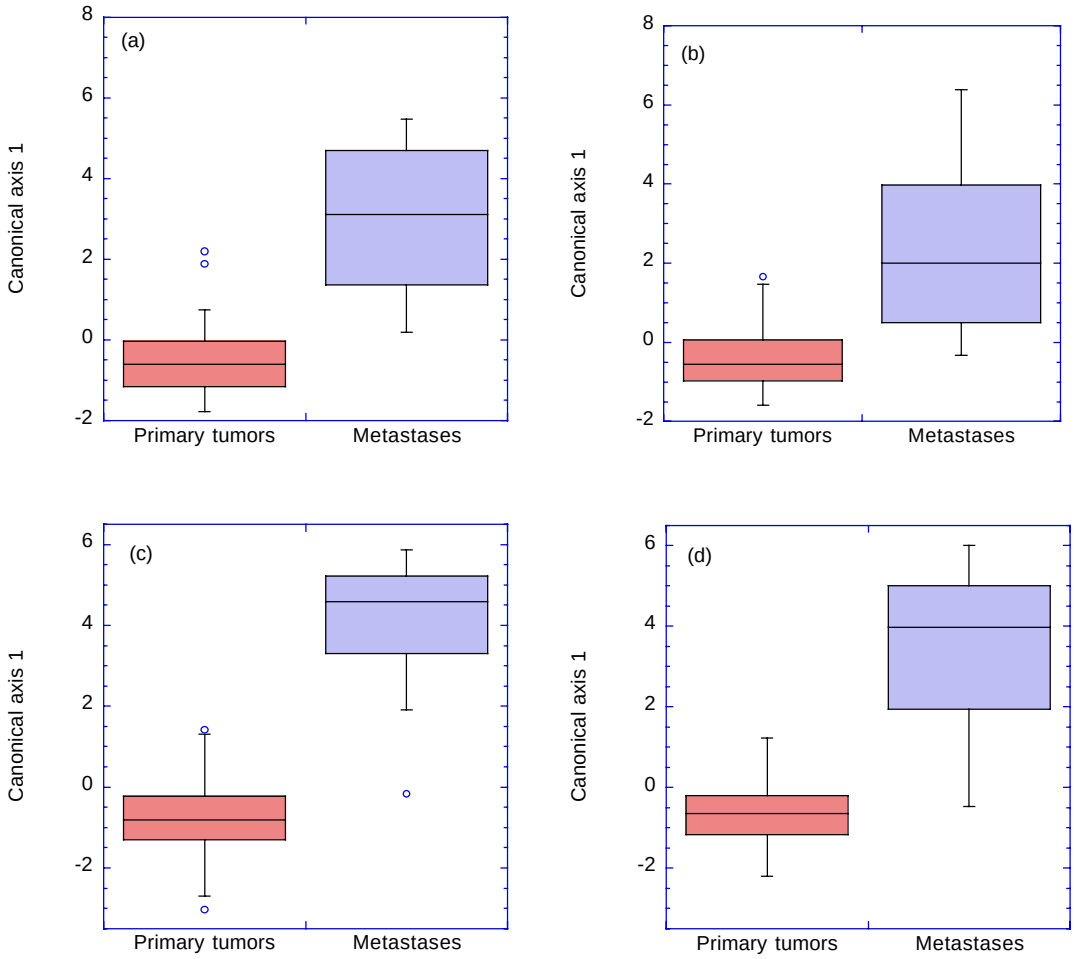


Fig. 5

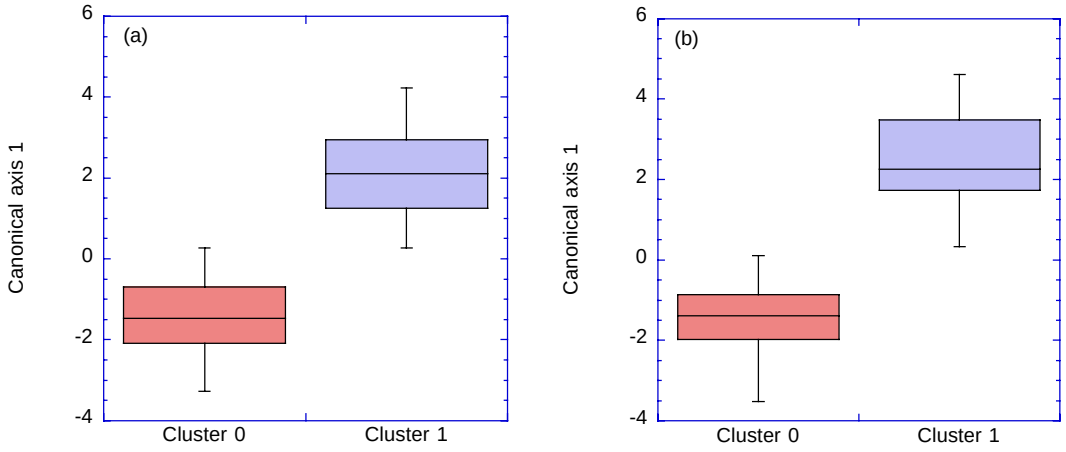

(a)
(b)
Canonical axis 1
Canonical axis 1
6
4
2
0
-2
-4
Cluster 0
Cluster 1
Cluster 0
Cluster 1

Fig. 6

UCRL-JRNL-203450

Supplement Report on Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The Special Case of Two Groups

J.R. Kercher , R.G. Langlois, B.A. Sokhansanj, C.F. Melius, J.N. Quong, F.P. Milanovich, B.W. Colston, Jr., K.W. Turteltaub, A.A. Quong

U.S. Department of Energy

Lawrence Livermore National Laboratory

Lawrence Livermore National Laboratory, Livermore, California 94551

This article was prepared for journal submission.

April 2004

DISCLAIMER

Supplement to Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The Special Case of Two Groups
PART A: Methods of Analysis

J. R.Kercher[1], R.G. Langlois[2], B.A. Sokhansanj[3],
C.F. Melius[4], J.N. Quong[4,5], F.P. Milanovich[6], B.W. Colston, Jr.[7],
K.W. Turteltaub[2], A.A. Quong[3,5]

[1]Environmental Sciences Division, L-235,
[2]Biodefense Division, [3]Chemical Biology and Nuclear Science Division,
4Chemistry and Chemical Engineering Division, [5]R Division
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, California 94551

[5]Georgetown University, Lombardi Cancer Center, 3970 Reservoir Road, NW,
Washington, DC 20057 USA

{NOTE TO THE EDITOR: If the main paper is accepted, we understand this supplemental report will be placed online.)

Section I. Supervised methods, the canonical equation, and Fisher's linear discriminant analysis

Partial list of supervised methods. The supervised methods include support vector machines (SVM, e.g., Furey et al. 2000, Xiong et al. 2001, Moler et al. 2000, Ambroise and McLachlan 2002), Fisher's linear discriminant function (or analysis) (FLDA, Xiong et al. 2000, Xiong et al. 2001, Dudoit et al. 2002), logistic regression (LR, Xiong et al. 2001, Kozak et al. 2003), linear predictor score (LPS,Wright et al 2003), maximum likelihood discriminant rules (ML, Dudoit et al. 2002, Nguyen and Rocke 2002), penalized discriminant analysis (PDA, Kari et al. 2003), nearest neighbor (Dudoit et al. 2002), partial least squares (PLS, Nguyen and Rocke 2002), classification and regression trees CART (Dudoit et al. 2002, Zhang et al. 2001), compound covariate predictor (Radmacher et al. 2002), polychotomous discrimination (Nguyen and Rocke 2002), and weighted voting (Golub et al. 1999).

Notation for the between-group variance-covariance matrix and the within-group variance-covariance matrix. The original data is represented by the matrix **X** whose elements x_{ij} are the value of the variable *i* of observation or experimental unit *j*. The index *i* ranges from 1 to *p* where *p* is the number of variables and the index *j* ranges from 1 to *N* where *N* is the number of observations. Let $\mathbf{x}_\mathbf{s}^j$ denote the *j*th column of **X**. The original data has been classified into *h* groups of observations. Denote the set of indices of the observations of group *k* by $J_k = \left\{ j \middle| \mathbf{x}_\mathbf{s}^j \textit{ is an observation in the kth group} \right\}$ where *k*=1,...,*h*. The number of observations in the *k*th group is n_k. The mean value of the variable *i* (protein concentration, gene expression ratio, etc) in the *k*th group is $\bar{x}_{i(k)} = (1/n_k) \sum_{j \in J_k} x_{ij}$ and the grand mean for the entire data set for the *i*th variable is given by $\bar{x}_i = (1/N) \sum_{j=1}^{N} x_{ij}$. The deviation of the *i*th variable of the *j*th observation from the grand mean is $t_{ij} = x_{ij} - \bar{x}_i$. By adding and subtracting the group means from the expression for t_{ij} and multiplying t_{ij} by itself and summing, we find a matrix equation $\mathbf{T} = \mathbf{W} + \mathbf{B} = \mathbf{X_W X_W^T} + \mathbf{X_B X_B^T}$, in which $T_{il} = \sum_{j=1}^{N} t_{ij} t_{lj}$, $W_{il} = \sum_{k=1}^{h} \sum_{j=1}^{n_k} \left(x_{ij} - \bar{x}_{i(k)} \right) \left(x_{lj} - \bar{x}_{l(k)} \right)$, $B_{il} = \sum_{k=1}^{h} n_k \left(\bar{x}_{i(k)} - \bar{x}_i \right) \left(\bar{x}_{l(k)} - \bar{x}_l \right)$, $\left(\mathbf{X_B} \right)_{ij} = \sqrt{n_k} \left(\bar{x}_{i(k)} - \bar{x}_i \right)$, and $\left(\mathbf{X_W} \right)_{ij} = \left(x_{ij} - \bar{x}_{i(k)} \right)$ where $j \in J_k$. The matrix **T** is the total sum-of-squares-and-cross-products (SSCP) matrix. The matrices **W** and **B** are the within-group and between-group SSCP matrices, respectively. The matrices **W** and **B** are related to the within-group and between-group variance-covariance matrices, respectively, by $\mathbf{\Omega} = \mathbf{W} / (N - h)$ and $\mathbf{\Theta} = \mathbf{B} / (h - 1)$, respectively. The degrees of freedom of **W** and **B** are $df_W = \sum_k (n_k - 1) = N - h$ and df_B=(h–1), respectively.

Assumptions of the canonical analysis. The vector $\mathbf{x}_s^j$ is the *j*th instance (observation) of the random variable **x**, which results from the model $\mathbf{x}=\boldsymbol{\mu}+\boldsymbol{\tau}_k+\boldsymbol{\varepsilon}$ where $\boldsymbol{\tau}_k$ is the effect of the observation coming from the *k*th population out of a total of *h* populations, $\boldsymbol{\mu}$ is the mean effect (grand mean of all populations), the vector $\boldsymbol{\varepsilon}$ is the error, and *j* belongs to the *k*th group, $j \in J_k$. Each of the vectors **x**, $\boldsymbol{\mu}$, $\boldsymbol{\tau}_k$, and $\boldsymbol{\varepsilon}$ have *p* components corresponding to the *p* measured variables. The vector $\boldsymbol{\varepsilon}$ is assumed to be independent and to be distributed in the *p* dimensions as $\boldsymbol{\varepsilon} \sim N_p(\mathbf{0},\boldsymbol{\Sigma})$ where N_p denotes the *p*-variate normal distribution. The null hypothesis for this model is $H_o: \tau_1 = \tau_2 = \cdots = \tau_h = 0$.

Maximizing the between-group dispersion. The canonical analysis procedure is to find the transformation from the old set of measured variables (protein concentrations, gene expression) to a new set (a linear combination of the old) such that in the new set the between-group variance-covariance matrix is maximized (Seal 1964). Let **E** be the matrix representing the transformation with matrix element E_{ij} and let $\mathbf{e}^i$ be the *i*th column vector of **E**. The variables in the new space are

$$y_{ij} = \sum_{k=1}^{p} x_{kj} E_{ki} \tag{S.1}$$

for the new variate *i* and observation *j*. In the new space the between-group dispersion along the *i*th canonical axis is

$$\Pi'_{ii} = \frac{1}{h-1} \sum_{m=1}^{r} \sum_{l=1}^{r} \sum_{k=1}^{h} \left(\bar{x}_{l(k)} - \bar{x}_l\right) E_{li} \left(\bar{x}_{m(k)} - \bar{x}_m\right) E_{mi} n_k \tag{S.2}$$

Taking the derivative with respect to E_{ij} to find the maximum dispersion produces an unbounded solution. To fix the scale to prevent unbounded solutions, we impose the condition that the within-group dispersion along each axis must be one. We use the Lagrange multiplier technique to do this, which leads to *p* equations to differentiate, one for each canonical axis.

$$\begin{aligned} D_i = \frac{1}{h-1} \sum_{m=1}^{r} \sum_{l=1}^{r} \sum_{k=1}^{h} \left(\bar{x}_{l(k)} - \bar{x}_l\right) E_{li} \left(\bar{x}_{m(k)} - \bar{x}_m\right) E_{mi} n_k \\ - \lambda'_i \left[\frac{1}{N-h} \sum_{lmkj} \left(x_{lj} - \bar{x}_{l(k)}\right) E_{li} \left(x_{mj} - \bar{x}_{m(k)}\right) E_{mi} - 1 \right] \end{aligned} \tag{S.3}$$

Differentiate eq. S.3 with respect to E_{ji}, set the result equal to zero to find the maximum, absorb the factors (*h*–1) and (*N*–*h*) into λ'_i and we find the equation

$$\sum_{mk}\left(\bar{x}_{j(k)}-\bar{x}_j\right)\left(\bar{x}_{m(k)}-\bar{x}_m\right)E_{mi}n_k-\lambda_i\sum_{lmk}\left(x_{jl}-\bar{x}_{j(k)}\right)\left(x_{ml}-\bar{x}_{m(k)}\right)E_{mi}=0 \qquad \text{(S.4)}$$

By holding *i* constant in eq. S.4 and varying *j* from 1 to *p*, we generate *p* equations, which we can write as one generalized-eigenvalue vector equation

$$(\mathbf{B}-\lambda_i\mathbf{W})\mathbf{e}_i=0 \qquad \text{(S.5)}$$

The canonical eigenvalue equation. There are *p* of these equations (eq. 13), one for each eigenvalue, and we can write all of them as a matrix equation

$$\mathbf{BE}-\mathbf{WE\Lambda}=0 \qquad \text{(S.6)}$$

where $\mathbf{\Lambda}$ is the matrix whose diagonal terms are eigenvalues λ_i and whose off-diagonal terms are zero.

Differentiating eq. S.3 with respect to λ_i produces a normalization equation.

$$\frac{1}{N-h}\mathbf{e}_i^{\mathrm{T}}\mathbf{W}\mathbf{e}_i=1 \qquad \text{(S.7)}$$

where the superscript T designates the transpose. This equation will be used to fix the "length" of the eigenvectors.

In the standard canonical analysis, eqs. S.5 and S.7 are solved and then eq. S.1 can be used to find the values of the observations in the new coordinates. Usually the t_{ij} values (eq. 3) are transformed to the new space rather than the x_{ij}. Up to this point we have not made any assumptions about the nature of **B** and **W**. If the number of degrees of freedom of the within-group variance (*N*–*h*) exceed the number of variables *p*, then **W** is non-singular and eq. S.5 can be solved by standard means. For example eq.S.5 converts to

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{e}_i=\lambda_i\mathbf{e} \qquad \text{(S.8)}$$

which is an eigenvalue equation for the matrix ($\mathbf{W}^{-1}\mathbf{B}$) and which can be solved like any other eigenvalue equation. Also, techniques exist to solve eq. S.5 directly if **B** and **W** are symmetric and **W** is positive definite.

However, in our case, because the number of variables greatly exceeds observations, **W** is singular, semi-positive definite, and $\mathbf{W}^{-1}$ does not exist. The singular nature of **W** rules out using eq. S.8, and we must use different approaches to analyze the canonical equation, eq. S.5.

Singular value decomposition of **W**. Index the singular values and singular vectors of **W** such that $\delta_i \geq \delta_j$ for *i*<*j*. Thus, $\delta_i = 0$ for *i*>*r* where *r* is the rank of **W**.

Equivalence of Fisher's linear discriminant analysis to standard canonical analysis for two groups. The transformation to the first canonical axis in canonical analysis is given by $y = \mathbf{e}^{1\mathbf{T}}\mathbf{x}_\mathbf{s}^i$ (eq. 1, Kercher et al. 2004) and the canonical equation for two groups is given by $\mathbf{We}' = \mathbf{s}^1$ (eq. 4, Kercher et al. 2004) where $\mathbf{e}^1 = c_1\mathbf{e}'$ and c_1 is a constant. From eq. S.17 below we have $\mathbf{s}^1 = c_2\left(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}\right)$ where c2 is a constant and $\bar{\mathbf{x}}_{(i)}$ is the group-mean vector for the *i*th group. Hence $y = c_1c_2\left(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}\right)^\mathbf{T}\mathbf{W}^{-1}\mathbf{x}_\mathbf{s}^i$. On the other hand Fisher's linear discriminant function transforms $\mathbf{x}_\mathbf{s}^i$ to a one-dimensional space. The transformation is $w = \left[\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}\right]^\mathbf{T}\mathbf{W}^{-1}\mathbf{x}_\mathbf{s}^i$ (Krzanowski 2000 p. 356). Except for an inconsequential constant scaling factor, the transformation for CA for two groups and Fisher's linear discriminant function are identical. The scaling factor is inconsequential because it factors out of the discrimination decision (e.g., Srivastava 2002 p. 247-248.)

Section II. Properties of Moore-Penrose generalized inverse (MPGI)

For the convenience of the reader and because it is a central result that we use to justify our use of the Moore-Penrose generalized inverse, we quote the definitions and theorem from Campbell and Meyer (1979, p. 28) regarding the minimal least squares solution to $\mathbf{We}' = \mathbf{s}^1$.

"*Definition*: … Suppose that $\mathbf{A} \in C^{m\times n}$ and $\mathbf{b} \in C^m$. Then a vector $\mathbf{u} \in C^n$ is called a least squares solution to **Ax**=**b** if $\|\mathbf{Au} - \mathbf{b}\| \leq \|\mathbf{Av} - \mathbf{b}\|$ for all $\mathbf{v} \in C^n$. A vector **u** is called a minimal least squares solution to **Ax**=**b** if **u** is a least squares solution to **Ax**=**b** and $\|\mathbf{u}\| < \|\mathbf{w}\|$ for all other least squares solutions **w**."

"*Theorem*. … Suppose that $\mathbf{A} \in C^{m\times n}$ and $\mathbf{b} \in C^m$. Then $\mathbf{A}^+\mathbf{b}$ is the minimal least squares solution to **Ax**=**b**."

Throughout the paper and supplement, we use properties of the MPGI regarding multiplying by a scalar, products of a matrix and MPGI of a matrix in various orders and combinations, MPGI of MPGI of a matrix, transpose of MPGI, etc as stated in standard texts. For example, these results are given in Schott (1997, Chap. 5, Sec. 2 through 7) and Campbell and Meyer (1979, Chap. 1 through 3).

Section III. An infinity of solutions.

Consider the generalized inverse approach in the case of two groups. From Kercher et al. (2004) we saw that the generalized inverse approximate solution to the canonical equation was given by

$$\mathbf{e}' = \mathbf{V}\mathbf{R}\mathbf{V}^{\mathbf{T}}\mathbf{s}^1 + \mathbf{V}[\mathbf{I} - \mathbf{R}\boldsymbol{\Delta}]\mathbf{V}^{\mathbf{T}}\mathbf{y} \qquad \text{(S.11)}$$

where **y** is any vector. In eq. 5 in Kercher et al. (2004) we chose the solution for which **y**=0. Note that while all solutions of the form of eq. S.11 give the same best approximate solution to eq. 4 in Kercher et al (2004), eq. 5 is the solution of minimal length.

Now because **WV**[**I**–**RΔ**]=**VΔ**[**I**–**RΔ**]=**0**, eq. 4 in Kercher et al (2004) remains unchanged no matter what the value of **y** is. Unfortunately, the original equation canonical equation, eq. 2 in Kercher et al (2004), which eq. 5 was derived from, is not invariant under changes in **y**. We first note that, without loss of generality, we may replace $\mathbf{V}^{\mathbf{T}}\mathbf{y}$ by the vector $\mathbf{y}' = \mathbf{V}^{\mathbf{T}}\mathbf{y}$ in eq. S.11. Next note that without loss of generality, we can assume that $\mathbf{y}'$ is of the form, $y_i' = 0$ for $i \le r$. Next we note that if we assume that $\left[\mathbf{V}^{\mathbf{T}}\mathbf{s}^1\right]_i \neq 0$, (which is the case in the example problems examined below), then we may define the matrix **Y** with matrix elements $Y_{ii} = y_i' \big/ \left[\mathbf{V}^{\mathbf{T}}\mathbf{s}^1\right]_i$ for $i>r$ and $Y_{ij}=0$ for all other matrix elements. That is, we may substitute $\mathbf{y} = \mathbf{V}\mathbf{y}' = \mathbf{V}\mathbf{Y}\mathbf{V}^{\mathbf{T}}\mathbf{s}^1$ in eq. S.11 without loss of generality and find that $\mathbf{e}'$ is of the form $\mathbf{e}' = \mathbf{V}\mathbf{R}\mathbf{V}^{\mathbf{T}}\mathbf{s}^1 + \mathbf{V}\mathbf{Y}\mathbf{V}^{\mathbf{T}}\mathbf{s}^1$ where **Y** is a diagonal matrix whose *r* upper-left diagonal entries are all zero. The lower-right (*p*–*r*) diagonal entries of **Y** can have any value.

Substituting the full solution for $\mathbf{e}'$ in eq. S.5 and taking the inner product with $\mathbf{e}'^{\mathbf{T}}$, we find the equation for λ to be

$$\lambda_1 = \frac{\xi_1}{\mathbf{s}^1\mathbf{V}\mathbf{R}\mathbf{V}^{\mathbf{T}}\mathbf{s}^1}\left[\sum_{i=1}^{r}\frac{\left(\mathbf{v}^i\cdot\mathbf{s}^1\right)^2}{\delta_i} + \sum_{i=r+1}^{p} Y_{ii}\left(\mathbf{v}^i\cdot\mathbf{s}^1\right)^2\right]^2.$$

We see immediately that λ_1 is bounded below by 0 for all values of Y_{ii}. If we restrict our family of solutions to values of $Y_{ii} \ge 0$, then λ_1 is bounded below by $\lambda_1 \ge \xi_1 \sum_{i=1}^{r}\left(\mathbf{v}_i\cdot\mathbf{s}^1\right)^2 \big/ \delta_i$ for $Y_{ii} \ge 0$. Hence the solution that we use in eq. 5 in Kercher et al (2004) is the most conservative of all solutions for which $Y_{ii} \ge 0$. By conservative, we mean producing the smallest value for the leading eigenvalue, which is proportional to the dispersion between the groups. Any solutions with all $Y_{ii} \ge 0$ corresponds, via the Moore-Penrose inverse, to a virtual within-group sum of squares matrix that is semi-positive definite; any solution, for which some $Y_{ii}<0$, corresponds to a virtual within-group sum of squares matrix that is not semi-positive definite. Therefore we attach more biological meaning to those solutions for which $Y_{ii} \ge 0$.

Section IV. Independence of the standard CA solutions on the scale

In the standard canonical analysis in which $p \le (N-h)$ and for which **W** is nonsingular, consider that the expressions for **W** and **B** can be written as $\mathbf{W} = \mathbf{X}_{\mathbf{W}}\mathbf{X}_{\mathbf{W}}^{\mathbf{T}}$ and $\mathbf{B} = \mathbf{X}_{\mathbf{B}}\mathbf{X}_{\mathbf{B}}^{\mathbf{T}}$ where the matrix elements of $\mathbf{X}_{\mathbf{W}}$ are given by $\left(\mathbf{X}_{\mathbf{W}}\right)_{ij} = x_{ij} - \bar{x}_{i(k)}$ and the

matrix elements of $\mathbf{X_B}$ are given by $(\mathbf{X_B})_{ik} = \sqrt{n_k}\left(\bar{x}_{i(k)} - \bar{x}_i\right)$ where *i* is the index for the variables, *j* is the index of observations, n_k is the number of observations in the *k*th group, and $j \in X_k$. A change of scale (or units) is obtained by multiplying each variable by a scale factor k_i and this is equivalent to pre-multiplying the $\mathbf{X_B}$ and $\mathbf{X_W}$ matrices by the square $p \times p$ diagonal matrix **K** where the diagonal values of **K** are given by the k_i. Thus in a new system of units, the **W** and **B** matrices become $\mathbf{W}' = \mathbf{K}\mathbf{X_W}\mathbf{X_W}^{\mathrm{T}}\mathbf{K}$ and $\mathbf{B}' = \mathbf{K}\mathbf{X_B}\mathbf{X_B}^{\mathrm{T}}\mathbf{K}$ and the canonical equation (eq. S.5) becomes $\left(\mathbf{W}'\right)^{-1}\mathbf{B}'\mathbf{e}' - \lambda'\mathbf{e}' = 0$ or $\mathbf{K}^{-1}\mathbf{W}^{-1}\mathbf{K}^{-1}\mathbf{K}\mathbf{B}\mathbf{K}\mathbf{e}' - \lambda'\mathbf{e}' = 0$, which simplifies to $\mathbf{W}^{-1}\mathbf{B}(\mathbf{K}\mathbf{e}') - \lambda'(\mathbf{K}\mathbf{e}') = 0$. We see immediately that if we set $\mathbf{e}^1 = \mathbf{K}\mathbf{e}'$ (or $\mathbf{e}' = \mathbf{K}^{-1}\mathbf{e}^1$) and $\lambda = \lambda'$ we recover the original canonical equation. Furthermore, the positions of the observations in canonical space remain unchanged after the change in units. That is, the vector form of eq. S.1 after a change in units becomes $\mathbf{y}'^{\mathrm{T}} = \mathbf{e}'^{1\mathrm{T}}\mathbf{X}' = \left(\mathbf{K}^{-1}\mathbf{e}^1\right)^{\mathrm{T}}\mathbf{K}\mathbf{X} = \mathbf{e}^{1\mathrm{T}}\mathbf{K}^{-1}\mathbf{K}\mathbf{X} = \mathbf{e}^{1\mathrm{T}}\mathbf{X} = \mathbf{y}^{\mathrm{T}}$ where the matrix **X** has matrix elements $(\mathbf{X})_{ij} = x_{ij} - \bar{x}_i$. Thus, we find that the positions of the observations are unchanged in canonical space in the standard canonical analysis.

This argument does not apply to the generalized inverse method for the singular **W** case. The preceding argument hinges on the property of true inverses that $(\mathbf{WK})^{-1} = \mathbf{K}^{-1}\mathbf{W}^{-1}$. However, in general, for the Moore-Penrose inverse, we have $(\mathbf{WK})^{+} \neq \mathbf{K}^{-1}\mathbf{W}^{+}$. More precisely, $(\mathbf{WK})^{+} = \mathbf{K}^{-1}\mathbf{W}^{+}$, if and only if, $\mathbf{W}^{+}\mathbf{W}\mathbf{K}\mathbf{K}^{\mathrm{T}}$ is symmetric (Schott 1997 Thm. 5.10). But while $\mathbf{W}^{+}\mathbf{W}$ and $\mathbf{K}^2$ are symmetric, $\mathbf{W}^{+}\mathbf{W}\mathbf{K}\mathbf{K}^{\mathrm{T}}$ is not. Hence we cannot use the preceding argument to show that the solutions are invariant to scale changes when **W** is singular.

Section V. Dependence of λ on K.

We note that in the preceding appendix we disproved a sufficient condition for invariance of the solutions of the generalized inverse to changes in scales, which is different from disproving a necessary condition. So, we shall find the explicit dependence of λ' on **K** in the generalized inverse approximate solution for the case of two groups. Define $\mathbf{W_K}$=**KWK**. Using the result $\mathbf{e}' = \mathbf{W_K}^{+}\mathbf{K}\mathbf{s}^1$ and $\mathbf{e}'$ is related to $\mathbf{e}'$ by a normalization constant, we find

$$\lambda = \frac{\mathbf{e}'^{\mathrm{T}}\mathbf{K}\mathbf{B}\mathbf{K}\mathbf{e}'}{\mathbf{e}'^{\mathrm{T}}\mathbf{K}\mathbf{W}\mathbf{K}\mathbf{e}'} = \frac{\mathbf{s}^{1\mathrm{T}}\mathbf{K}\mathbf{W_K}^{+}\mathbf{K}\mathbf{B}\mathbf{K}\mathbf{W_K}^{+}\mathbf{K}\mathbf{s}^1}{\mathbf{s}^{1\mathrm{T}}\mathbf{K}\mathbf{W_K}^{+}\mathbf{W_K}\mathbf{W_K}^{+}\mathbf{K}\mathbf{s}^1} = \xi_1 \frac{\mathbf{s}^{1\mathrm{T}}\mathbf{K}\mathbf{W_K}^{+}\mathbf{K}\mathbf{s}^1\left(\mathbf{s}^{1\mathrm{T}}\mathbf{K}\mathbf{W_K}^{+}\mathbf{K}\mathbf{s}^1\right)}{\mathbf{s}^{1\mathrm{T}}\mathbf{K}\mathbf{W_K}^{+}\mathbf{K}\mathbf{s}^1} \qquad \text{(S.12)}$$
$$= \xi_1 \mathbf{s}^{1\mathrm{T}}\mathbf{K}\mathbf{W_K}^{+}\mathbf{K}\mathbf{s}^1$$

We can differentiate eq. S.12 with respect to k_i, the *i*th diagonal term of **K**. This leads to

$$\frac{\partial \lambda'}{\partial k_i} = 2\xi_1 \mathbf{s}^{1\mathbf{T}} \mathbf{K}\left[\mathbf{I} - \mathbf{W}_\mathbf{K}^+ \mathbf{W}_\mathbf{K}\right] \mathbf{D}_i \mathbf{K}^{-1} \mathbf{W}_\mathbf{K}^+ \mathbf{K}\mathbf{s}^1 + 2\xi_1 \mathbf{s}^{1\mathbf{T}} \mathbf{K}\mathbf{W}_\mathbf{K}^+ \mathbf{K}^{-1} \mathbf{D}_i \left[\mathbf{I} - \mathbf{W}_\mathbf{K}^+ \mathbf{W}_\mathbf{K}\right] \mathbf{K}\mathbf{s}^1 \qquad \text{(S.13)}$$

We see that if **W** is nonsingular, then $\mathbf{W}_\mathbf{K}$ is nonsingular, which implies that $\mathbf{W}_\mathbf{K}^+ = \mathbf{W}_\mathbf{K}^{-1}$, and therefore from eq. S.13, $\frac{\partial \lambda'}{\partial k_i} = 0$. Otherwise, for the Moore-Penrose generalized inverse, which we use when $\mathbf{W}_\mathbf{K}$ is singular, eq. S.13 implies $\frac{\partial \lambda'}{\partial k_i} \neq 0$.

Note that if all diagonal elements of **K** are equal to each other, say **K**=*k***I**, then the Moore-Penrose inverse has the property $(\mathbf{KWK})^+ = \left(k^2 \mathbf{IWI}\right)^+ = k^{-2}\mathbf{W}^+$. We find that λ is invariant under uniform scale changes for all variables, **K**=*k***I**, i.e.,

$$\lambda' = \frac{\mathbf{e}'^{\mathbf{T}} \mathbf{KBK}\mathbf{e}'}{\mathbf{e}'^{\mathbf{T}} \mathbf{KWK}\mathbf{e}'} = \frac{\mathbf{s}^{1\mathbf{T}} \mathbf{K}\mathbf{W}_\mathbf{K}^+ k\mathbf{I}\mathbf{B}k\mathbf{I}\mathbf{W}_\mathbf{K}^+ \mathbf{K}\mathbf{s}^1}{\mathbf{s}^{1\mathbf{T}} \mathbf{K}\mathbf{W}_\mathbf{K}^+ k\mathbf{I}\mathbf{W}k\mathbf{I}\mathbf{W}_\mathbf{K}^+ \mathbf{K}\mathbf{s}^1} = \frac{\mathbf{s}^{1\mathbf{T}} k\mathbf{I}k^{-2}\mathbf{W}^+\mathbf{B}k^{-2}\mathbf{W}^+ k\mathbf{I}\mathbf{s}^1}{\mathbf{s}^{1\mathbf{T}} k\mathbf{I}k^{-2}\mathbf{W}^+\mathbf{W}k^{-2}\mathbf{W}^+ k\mathbf{I}\mathbf{s}^1}$$
$$= \frac{\mathbf{s}^{1\mathbf{T}} \mathbf{W}^+\mathbf{B}\mathbf{W}^+\mathbf{s}^1}{\mathbf{s}^{1\mathbf{T}} \mathbf{W}^+\mathbf{W}\mathbf{W}^+\mathbf{s}^1} = \frac{\mathbf{e}^{1\mathbf{T}} \mathbf{B}\mathbf{e}^1}{\mathbf{e}^{1\mathbf{T}} \mathbf{W}\mathbf{e}^1} = \lambda$$

This same result can also be obtained from eq. S.13 by noting that for **K**=*k***I**, then $\mathbf{D}_i \to \mathbf{I}$ when differentiating λ with respect to *k*. Then using $\left[\mathbf{I} - \mathbf{W}_\mathbf{K}^+ \mathbf{W}_\mathbf{K}\right] \mathbf{W}_\mathbf{K} = 0$ and the fact that $\mathbf{W}_\mathbf{K}$ commutes with $\mathbf{W}_\mathbf{K}^+$ because $\mathbf{W}_\mathbf{K}$ is symmetric, eq. A.5 immediately implies that λ does not depend on *k*.

Section VI. Explicit dependence of group separation on scales and variables.

We saw that in the standard canonical analysis for **W** nonsingular, the positions of each sample in canonical space are invariant under change in the units or scales in which the variables are measured, i.e., $\mathbf{y}' = \mathbf{y}$. Now consider the separation of groups in the two-group case using the generalized inverse approach when the number of variables exceeds the number of observations. Let **L** be the vector difference of the means of the two groups in the space of the original measurement variables. So the components L_i of **L** are given by $L_i = \frac{1}{n_1}\sum_{j \in J_1} x_{ij} - \frac{1}{n_2}\sum_{j \in J_2} x_{ij}$. Let **K** be a scale factor matrix in which all diagonal elements are equal to 1 except for the *i*th diagonal element, which is equal to (*k*+1), i.e., $\mathbf{K} = \mathbf{I} + k\mathbf{d}_i\mathbf{d}_i^\mathbf{T}$ where $\mathbf{d}_i$ is the unit vector in the *i*th direction or *i*th variable.

Note that for all this section the *i*th direction in measured-variable space is a special direction in which the scale has been multiplied by *k*. We show in the next section that $\mathbf{s}^1 = c\mathbf{L}$ where *c* is a normalization constant. The contribution of the *j*th variable to the separation of the two groups in transformed canonical space is given by

$\mathrm{T}_j = \mathbf{e}^{\mathrm{T}}\mathbf{d}_j L_j \propto \left(\mathbf{s}^{\mathrm{T}}\mathbf{K}\mathbf{W}_{\mathbf{K}}^{+}\right)\left(\mathbf{K}\mathbf{d}_j L_j\right) \propto \mathbf{L}^{\mathrm{T}}\mathbf{K}\mathbf{W}_{\mathbf{K}}^{+}\mathbf{K}\mathbf{d}_j L_j$. Then the ratio of the contribution of the *i*th variable to the total separation of the two groups is given by

$$G_i = \frac{\mathrm{T}_i}{\mathrm{T}} = \frac{\mathbf{L}^{\mathrm{T}}\mathbf{K}\mathbf{W}_{\mathbf{K}}^{+}\mathbf{K}\mathbf{d}_i L_i}{\mathbf{L}^{\mathrm{T}}\mathbf{K}\mathbf{W}_{\mathbf{K}}^{+}\mathbf{K}\mathbf{L}_{\perp} + \mathbf{L}^{\mathrm{T}}\mathbf{K}\mathbf{W}_{\mathbf{K}}^{+}\mathbf{K}\mathbf{d}_i L_i} = \frac{\mathbf{L}^{\mathrm{T}}\mathbf{K}\mathbf{W}_{\mathbf{K}}^{+}\mathbf{K}\mathbf{d}_i L_i}{\mathbf{L}^{\mathrm{T}}\mathbf{K}\mathbf{W}_{\mathbf{K}}^{+}\mathbf{K}\mathbf{L}} \text{ where } \mathbf{L}_{\perp} = \sum_{j \neq i} L_j \mathbf{d}_j \text{ and}$$

$\mathbf{L} = \mathbf{L}_{\perp} + L_i \mathbf{d}_i$

To develop an explicit expression for the dependence of G_i on *k*, we will use a representation of the Moore-Penrose inverse developed by Meyer (1973) for modified matrices. See Campbell and Meyer (1979, Thm 3.1.3) for applications. The matrix $\mathbf{W}_{\mathbf{K}}$=**KWK** can be written in the form

$$\mathbf{W}_{\mathbf{K}} = \mathbf{KWK} = \begin{pmatrix} w_{11} & \bullet & w_{1i}(k+1) & \bullet & w_{1p} \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ (k+1)w_{i1} & \bullet & (k+1)^2 w_{ii} & \bullet & (k+1)w_{ip} \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ w_{p1} & \bullet & w_{pi}(k+1) & \bullet & w_{pp} \end{pmatrix}$$

$$= \begin{pmatrix} w_{11} & \bullet & w_{1i} & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ w_{i1} & \bullet & w_{ii} & \bullet & w_{ip} \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ w_{p1} & \bullet & w_{pi} & \bullet & w_{pp} \end{pmatrix} + \begin{pmatrix} 0 & \cdots & kw_{1i} & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ kw_{i1} & \cdots & (2k+k^2)w_{ii} & \cdots & kw_{ir} \\ 0 & \cdots & \vdots & 0 & \cdots \\ \vdots & \ddots & kw_{ri} & \vdots & \ddots \end{pmatrix}$$

$$= \mathbf{W} + \mathbf{\Phi}$$

The second matrix $\mathbf{\Phi}$ is of rank 2, which can be seen by finding that there are only two nonzero eigenvalues. We will make the following decomposition of **KWK**

$$\mathbf{W_K} = \mathbf{KWK} = \mathbf{W} + k\begin{pmatrix} 0 & \cdots & w_{1i} & 0 & \cdots \\ 0 & \cdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & w_{ii} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & w_{ri} & 0 & \cdots \end{pmatrix} + k\begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{i1} & \cdots & (k+1)w_{ii} & \cdots & w_{ir} \\ 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \end{pmatrix}$$

$$= \mathbf{W} + k\begin{pmatrix} w_{1i} \\ \vdots \\ w_{ii} \\ \vdots \\ w_{ri} \end{pmatrix}\begin{pmatrix} 0 & \cdots & 1 & 0 & \cdots \end{pmatrix} + k\begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{pmatrix}\begin{pmatrix} w_{i1} & \cdots & [k+1]w_{ii} & \cdots & w_{ir} \end{pmatrix}$$

$$= \mathbf{W} + \mathbf{w}_i\left(k\mathbf{d}_i^{\mathbf{T}}\right) + \left(k\mathbf{d}_i\right)\mathbf{z}_i^{\mathbf{T}}$$

where the vector $\mathbf{w}_i$ is given by $\mathbf{w}_i = \begin{pmatrix} w_{i1} & \cdots & w_{ii} & \cdots & w_{ir} \end{pmatrix}^{\mathbf{T}}$ and the vector $\mathbf{z}_i$ is given by $\mathbf{z}_i = \mathbf{w}_i + kw_{ii}\mathbf{d}_i$. We can write $\mathbf{P} = \mathbf{W} + \mathbf{w}_i\left(k\mathbf{d}_i^{\mathbf{T}}\right)$ and then note that $\mathbf{w}_i$ is in the range of $\mathbf{W}$ $\left(\mathbf{w}_i \in R(\mathbf{W})\right)$ and $\mathbf{z}_i$ is in the range of $\mathbf{P}^{\mathbf{T}}$. These observations indicate that we should use Thm 3 and Thm 5 from Meyer (1973) for the first and second augmentation calculation, respectively. That is we apply Thm 3 and Thm 5 iteratively, so that

$$\mathbf{P}^+ = \mathbf{W}^+ + \frac{k\left(I - \mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i\mathbf{w}_i^{\mathbf{T}}\mathbf{W}^+\mathbf{W}^+}{1 + k\mathbf{d}_i^{\mathbf{T}}\mathbf{W}^+\mathbf{w}_i}$$
$$- \frac{\left[\left(\mathbf{w}_i^{\mathbf{T}}\mathbf{W}^+\mathbf{W}^+\mathbf{w}_i\right)k\left(\mathbf{I} - \mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i + \left(1 + k\mathbf{d}_i^{\mathbf{T}}\mathbf{W}^+\mathbf{w}_i\right)\mathbf{W}^+\mathbf{w}_i\right]}{\left(\mathbf{w}_i^{\mathbf{T}}\mathbf{W}^+\mathbf{W}^+\mathbf{w}_i\right)k^2\mathbf{d}_i^{\mathbf{T}}\left(\mathbf{I} - \mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i + \left(1 + k\mathbf{d}_i^{\mathbf{T}}\mathbf{W}^+\mathbf{w}_i\right)^2} \times \qquad \text{(S.14a)}$$
$$\left[\frac{k^2\mathbf{d}_i^{\mathbf{T}}\left(\mathbf{I} - \mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i}{1 + k\mathbf{d}_i^{\mathbf{T}}\mathbf{W}^+\mathbf{w}_i}\mathbf{w}_i^{\mathbf{T}}\mathbf{W}^+\mathbf{W}^+ + k\mathbf{d}_i^{\mathbf{T}}\mathbf{W}^+\right]$$

and

$$\mathbf{W}_{\mathbf{K}}^+ = \left(\mathbf{KWK}\right)^+ = \mathbf{P}^+ + \frac{k\mathbf{P}^+\left(\mathbf{P}^+\right)^{\mathbf{T}}\mathbf{z}_i\mathbf{d}_i^{\mathbf{T}}\left(\mathbf{I} - \mathbf{PP}^+\right)}{1 + k\mathbf{z}_i^{\mathbf{T}}\mathbf{P}^+\mathbf{d}_i}$$
$$- \frac{\left[k^2\left\{\mathbf{d}_i^{\mathbf{T}}\left(\mathbf{I} - \mathbf{PP}^+\right)\mathbf{d}_i\right\}\mathbf{P}^+\left(\mathbf{P}^+\right)^{\mathbf{T}}\mathbf{z}_i + k\left(1 + k\mathbf{z}_i^{\mathbf{T}}\mathbf{P}^+\mathbf{d}_i\right)\mathbf{P}^+\mathbf{d}_i\right]}{\left[k^2\left\{\mathbf{z}_i^{\mathbf{T}}\mathbf{P}^+\left(\mathbf{P}^+\right)^{\mathbf{T}}\mathbf{z}_i\right\}\mathbf{d}_i^{\mathbf{T}}\left(\mathbf{I} - \mathbf{PP}^+\right)\mathbf{d}_i + \left(1 + k\mathbf{z}_i^{\mathbf{T}}\mathbf{P}^+\mathbf{d}_i\right)^2\right]} \times \qquad \text{(S.14b)}$$
$$\left[\frac{\mathbf{z}_i^{\mathbf{T}}\mathbf{P}^+\left(\mathbf{P}^+\right)^{\mathbf{T}}\mathbf{z}_i}{1 + k\mathbf{z}_i^{\mathbf{T}}\mathbf{P}^+\mathbf{d}_i}k\mathbf{d}_i^{\mathbf{T}}\left(\mathbf{I} - \mathbf{PP}^+\right) + \mathbf{z}_i^{\mathbf{T}}\mathbf{P}^+\right]$$

While the formulae for $\mathbf{P}^+$ and $(\mathbf{KWK})^+$ are cumbersome, they possess the important characteristic that they are well-behaved for small *k*. In fact, one can see by inspection that both $(\mathbf{KWK})^+ \to \mathbf{P}^+$ and $\mathbf{P}^+ \to \mathbf{W}^+$, continuously, as $k \to 0$. Note that $\mathbf{W}^+ = \mathbf{W}^+\mathbf{W}\mathbf{W}^+ = \mathbf{W}^+\mathbf{W}^+\mathbf{W} = \left(\mathbf{W}^+\right)^2\mathbf{W}$ and thus $\mathbf{u}_j = \left(\mathbf{W}^+\right)^2\mathbf{w}_j$ where $\mathbf{u}_j$ is the *j*th column of the matrix $\mathbf{W}^+$. Also it can be shown that $\mathbf{w}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{w}_i = w_{ii}$, $\mathbf{W}\mathbf{W}^+\mathbf{w}_i = \mathbf{w}_i$, $\mathbf{P}\mathbf{P}^+ = \mathbf{W}\mathbf{W}^+ = \mathbf{W}^+\mathbf{W}$, and $\mathbf{z}_i^{\mathrm{T}}\mathbf{P}^+ = \mathbf{w}_i^{\mathrm{T}}\mathbf{W}^+$. Using these relations we can simplify the parts of the numerator and denominator of G_i to

$$\mathbf{L}^{\mathrm{T}}\mathbf{K}\mathbf{W}_{\mathbf{K}}^+\mathbf{K}\mathbf{d}_i L_i = \frac{(1+k)\left(\mathbf{L}_\perp^{\mathrm{T}}\mathbf{P}^+ + (1+k)L_i\mathbf{d}_i^{\mathrm{T}}\mathbf{P}^+\right)}{\mathbf{w}_i^{\mathrm{T}}\mathbf{u}_i k^2\mathbf{d}_i^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i + \left(1+k\mathbf{w}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{d}_i\right)^2} \times \left[\mathbf{W}^+\mathbf{w}_i k\mathbf{d}_i^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^+\right)\mathbf{d}_i + \mathbf{d}_i\left(1+k\mathbf{w}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{d}_i\right)\right]L_i \tag{S.15}$$

and

$$\begin{aligned}\mathbf{L}^{\mathrm{T}}\mathbf{K}\mathbf{W}_{\mathbf{K}}^+\mathbf{K}\mathbf{L}_\perp &= \left[\mathbf{L}_\perp^{\mathrm{T}} + (1+k)L_i\mathbf{d}_i^{\mathrm{T}}\right]\mathbf{P}^+\mathbf{L}_\perp + \\ &\frac{\left[\left(\mathbf{L}_\perp^{\mathrm{T}} + (1+k)L_i\mathbf{d}_i^{\mathrm{T}}\right)\mathbf{P}^+\mathbf{W}^+\mathbf{w}_i\right]k\left[\mathbf{d}_i^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{L}_\perp\right]}{1+k\mathbf{w}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{d}_i} \\ &- \frac{\left[\dfrac{k^2\mathbf{d}_i^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i}{1+k\mathbf{w}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{d}_i}\left(\mathbf{L}_\perp^{\mathrm{T}} + (1+k)L_i\mathbf{d}_i^{\mathrm{T}}\right)\mathbf{P}^+\mathbf{W}^+\mathbf{w}_i + k\left(\mathbf{L}_\perp^{\mathrm{T}} + (1+k)L_i\mathbf{d}_i^{\mathrm{T}}\right)\mathbf{P}^+\mathbf{d}_i\right]}{\left(\mathbf{w}_i^{\mathrm{T}}\mathbf{u}_i\right)k^2\left[\mathbf{d}_i^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i\right] + \left(1+k\mathbf{w}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{d}_i\right)^2} \times \\ &\left\{\left(\mathbf{w}_i^{\mathrm{T}}\mathbf{u}_i\right)k\left[\mathbf{d}_i^{\mathrm{T}}\left(1-\mathbf{W}^+\mathbf{W}\right)\mathbf{L}_\perp\right] + \left(1+k\mathbf{w}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{d}_i\right)\mathbf{w}_\mathbf{i}^{\mathrm{T}}\mathbf{W}^+\mathbf{L}_\perp\right\}\end{aligned} \tag{S.16}$$

where $\mathbf{d}_i^{\mathrm{T}}\mathbf{P}^+ = \dfrac{k\left[\mathbf{d}_i^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i\right]\mathbf{u}_i^{\mathrm{T}} + \left(1+k\mathbf{d}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{w}_i\right)\mathbf{d}_i^{\mathrm{T}}\mathbf{W}^+}{\left(\mathbf{w}_i^{\mathrm{T}}\mathbf{u}_i\right)k^2\left[\mathbf{d}_i^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i\right] + \left(1+k\mathbf{d}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{w}_i\right)^2}$, and

$$\begin{aligned}\mathbf{L}_\perp^{\mathrm{T}}\mathbf{P}^+ &= \mathbf{L}_\perp^{\mathrm{T}}\mathbf{W}^+ + \frac{k\left[\mathbf{L}_\perp^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i\right]\mathbf{u}_i^{\mathrm{T}}}{1+k\left(\mathbf{d}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{w}_i\right)} \\ &- \frac{\left\{\left(\mathbf{w}_i^{\mathrm{T}}\mathbf{u}_i\right)k\left[\mathbf{L}_\perp^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i\right] + \left(1+k\mathbf{d}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{w}_i\right)\left(\mathbf{L}_\perp^{\mathrm{T}}\mathbf{W}^+\mathbf{w}_i\right)\right\}}{\left(\mathbf{w}_i^{\mathrm{T}}\mathbf{u}_i\right)k^2\left[\mathbf{d}_i^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i\right] + \left(1+k\mathbf{d}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{w}_i\right)^2} \times \\ &\left\{\frac{k^2\left[\mathbf{d}_i^{\mathrm{T}}\left(\mathbf{I}-\mathbf{W}^+\mathbf{W}\right)\mathbf{d}_i\right]}{1+k\mathbf{d}_i^{\mathrm{T}}\mathbf{W}^+\mathbf{w}_i}\mathbf{u}_i^{\mathrm{T}} + k\mathbf{d}_i^{\mathrm{T}}\mathbf{W}^+\right\}\end{aligned}$$

Section VII. Contribution of the separation of two groups from variable *i*.

For the case of two groups we can find the explicit formula for the contribution of variable *i* to the total distance in canonical space between the two groups. As argued by Kercher et al (2004) the canonical equation for two groups is given by

$$\mathbf{We'} = \mathbf{s}^1$$

where **W** is the within group sum of squares and cross products matrix and $\mathbf{s}^1$ is the eigenvector of **B** with nonzero eigenvalue. First find an explicit formula for the eigenvector of **B**. Note that **B** is given by

$$\mathbf{B} = \mathbf{X_B X_B^T}$$

where $\mathbf{X_B}$ is given by

$$\mathbf{X_B} = \begin{pmatrix} \sqrt{n_1}\left(\bar{x}_{1(1)} - \bar{x}_1\right) & \sqrt{n_2}\left(\bar{x}_{1(2)} - \bar{x}_1\right) \\ \sqrt{n_1}\left(\bar{x}_{2(1)} - \bar{x}_2\right) & \sqrt{n_2}\left(\bar{x}_{2(2)} - \bar{x}_2\right) \\ \vdots & \vdots \\ \sqrt{n_1}\left(\bar{x}_{p(1)} - \bar{x}_p\right) & \sqrt{n_2}\left(\bar{x}_{p(2)} - \bar{x}_p\right) \end{pmatrix}$$

where $\bar{x}_{i(j)}$ is the group mean for the *i*th variable for the *j*th group and $\bar{x}_i$ is the global mean for the *i*th variable, n_j is the number of members of the *j*th group. We can rewrite $\mathbf{X_B}$ as

$$\mathbf{X_B} = \frac{n_1 n_2}{n_1 + n_2} \begin{pmatrix} \frac{\left(\bar{x}_{1(1)} - \bar{x}_{1(2)}\right)}{\sqrt{n_1}} & -\frac{\left(\bar{x}_{1(1)} - \bar{x}_{1(2)}\right)}{\sqrt{n_2}} \\ \frac{\left(\bar{x}_{2(1)} - \bar{x}_{2(2)}\right)}{\sqrt{n_1}} & -\frac{\left(\bar{x}_{2(1)} - \bar{x}_{2(2)}\right)}{\sqrt{n_2}} \\ \vdots & \vdots \\ \frac{\left(\bar{x}_{p(1)} - \bar{x}_{p(2)}\right)}{\sqrt{n_1}} & -\frac{\left(\bar{x}_{p(1)} - \bar{x}_{p(2)}\right)}{\sqrt{n_2}} \end{pmatrix}$$

Consider any vector **a** with components a_i. Then

$$\mathbf{X_B^T a} = \frac{n_1 n_2}{n_1 + n_2} \begin{pmatrix} \frac{\left(\bar{x}_{1(1)} - \bar{x}_{1(2)}\right)}{\sqrt{n_1}} & \frac{\left(\bar{x}_{2(1)} - \bar{x}_{2(2)}\right)}{\sqrt{n_1}} & \cdots & \frac{\left(\bar{x}_{p(1)} - \bar{x}_{p(2)}\right)}{\sqrt{n_1}} \\ -\frac{\left(\bar{x}_{1(1)} - \bar{x}_{1(2)}\right)}{\sqrt{n_2}} & -\frac{\left(\bar{x}_{2(1)} - \bar{x}_{2(2)}\right)}{\sqrt{n_2}} & \cdots & -\frac{\left(\bar{x}_{p(1)} - \bar{x}_{p(2)}\right)}{\sqrt{n_2}} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$$

or

$$\mathbf{X_B^T a} = \frac{n_1 n_2}{n_1 + n_2}\begin{pmatrix} \frac{1}{\sqrt{n_1}}\sum_{i=1}^{p}\left(x_{i(1)} - x_{i(2)}\right)a_i \\ \frac{-1}{\sqrt{n_2}}\sum_{i=1}^{p}\left(x_{i(1)} - x_{i(2)}\right)a_i \end{pmatrix} = \frac{\psi\sqrt{n_1 n_2}}{n_1 + n_2}\begin{pmatrix} \sqrt{n_2} \\ -\sqrt{n_1} \end{pmatrix}$$

where

$$\psi = \sum_{i=1}^{p}\left(\bar{x}_{i(1)} - \bar{x}_{i(2)}\right)a_i$$

Therefore

$$\mathbf{X_B X_B^T a} = \frac{n_1 n_2}{n_1 + n_2}\begin{pmatrix} \frac{\left(\bar{x}_{1(1)} - \bar{x}_{1(2)}\right)}{\sqrt{n_1}} & -\frac{\left(\bar{x}_{1(1)} - \bar{x}_{1(2)}\right)}{\sqrt{n_2}} \\ \frac{\left(\bar{x}_{2(1)} - \bar{x}_{2(2)}\right)}{\sqrt{n_1}} & -\frac{\left(\bar{x}_{2(1)} - \bar{x}_{2(2)}\right)}{\sqrt{n_2}} \\ \vdots & \vdots \\ \frac{\left(\bar{x}_{p(1)} - \bar{x}_{p(2)}\right)}{\sqrt{n_1}} & -\frac{\left(\bar{x}_{p(1)} - \bar{x}_{p(2)}\right)}{\sqrt{n_2}} \end{pmatrix} \frac{\psi\sqrt{n_1 n_2}}{n_1 + n_2}\begin{pmatrix} \sqrt{n_2} \\ -\sqrt{n_1} \end{pmatrix}$$

or

$$\mathbf{X_B X_B^T a} = \frac{\psi n_1 n_2}{\left(n_1 + n_2\right)}\begin{pmatrix} \left(\bar{x}_{1(1)} - \bar{x}_{1(2)}\right) \\ \left(\bar{x}_{2(1)} - \bar{x}_{2(2)}\right) \\ \vdots \\ \left(\bar{x}_{p(1)} - \bar{x}_{p(2)}\right) \end{pmatrix}$$

Hence **B** transforms any vector to a vector proportional to $\mathbf{s}^1$ where $\mathbf{s}^1$ is

$$\mathbf{s}^1 = c\begin{pmatrix} \left(\bar{x}_{1(1)} - \bar{x}_{1(2)}\right) \\ \left(\bar{x}_{2(1)} - \bar{x}_{2(2)}\right) \\ \vdots \\ \left(\bar{x}_{p(1)} - \bar{x}_{p(2)}\right) \end{pmatrix} = c_2\left[\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}\right] \qquad \text{(S.17)}$$

where c_2 is a normalization constant. Now the contribution to the distance between the means that is due to variable *i* is proportional to $e_i'\left(\bar{x}_{i(1)} - \bar{x}_{i(2)}\right)$ where e_i' is the *i*th

component of $\mathbf{e}'$, the solution to $\mathbf{W}\mathbf{e}' = \mathbf{s}^1$. Then the fraction of the distance between the means that is due to variable *i* is given by

$$G_i(0) = \frac{e_i'\left(\bar{x}_{i(1)} - \bar{x}_{i(2)}\right)}{\sum_{j=1}^{p} e_j'\left(\bar{x}_{j(1)} - \bar{x}_{j(2)}\right)}$$

Let **L** be the vector with components $L_j = \left(\bar{x}_{j(1)} - \bar{x}_{j(2)}\right)$ and we select the minimal length least squares solution of eq. S.11 so that

$$\mathbf{e}' = \mathbf{W}^{+}\mathbf{s}^1 = c\mathbf{W}^{+}\mathbf{L}$$

So $G_i(0)$ is given by

$$G_i(0) = \frac{L_i \mathbf{d}_i^{\mathbf{T}} \mathbf{W}^{+}\mathbf{L}}{\mathbf{L}^{\mathbf{T}}\mathbf{W}^{+}\mathbf{L}}$$

where $\mathbf{d}_i$ is defined in the previous section. Note that $G_i(0)$ is the fraction of the total generalized Mahalanobis distance that is due to component *i* or variable *i*.

Section VIII. Filtering Generalized Inverse Canonical Analysis Results for Variable Selection in the Backward Elimination Algorithm

Criterion A: Sensitivity of between-group SSCP. In this criterion we assume that those variables to which the between group variance in canonical space is most sensitive are the most important. Calculate $\Lambda_{orig} = \det\left|\mathbf{W}_{orig}\right| / \det\left|\mathbf{W}_{orig} + \mathbf{B}_{orig}\right|$ where $\left(\mathbf{W}_{orig}\right)_{lm} = \sum_{s,o=1}^{p}\sum_{j=1}^{N} E_{ls}^{T}\left(x_{sj} - \bar{x}_{s(k)}\right)E_{mo}^{T}\left(x_{oj} - \bar{x}_{o(k)}\right)$ and $\left(\mathbf{B}_{orig}\right)_{lm} = \sum_{s,o=1}^{p}\sum_{k=1}^{h} E_{ls}^{T}\left(\bar{x}_{s(k)} - \bar{x}_{s}\right)E_{mo}^{T}\left(\bar{x}_{o(k)} - \bar{x}_{o}\right)$. Then selecting variable *i* set $E_{il}' = 1.1E_{il}$ for all *l* from 1 to (*h*–1) and $E_{jl}' = E_{jl}$ for all *l* and for $j \neq i$. Then define $\Lambda_{new} = \det\left|\mathbf{W}_{orig}\right| / \det\left|\mathbf{W}_{orig} + \mathbf{B}_{new}\right|$ where $\left(\mathbf{B}_{new}\right)_{lm} = \sum_{s,o}\sum_{k} E_{sl}'\left(\bar{x}_{s(k)} - \bar{x}_{s}\right)E_{om}'\left(\bar{x}_{o(k)} - \bar{x}_{o}\right)$. Define the sensitivity to the variable *i* as $\Gamma_i = \left[\left(\Lambda_{orig} - \Lambda_{new}\right)/\Lambda_{orig}\right] \Big/ \left[\left(\sum_l E_{il}' - \sum_l E_{il}\right) \Big/ \sum_l E_{il}\right] = 10 \cdot \left(\Lambda_{orig} - \Lambda_{new}\right)/\Lambda_{orig}$. After calculating the sensitivities of all the variables, we sort on the sensitivities and retain only the (*N*–*h*) largest. These variables are used in the stepwise procedure. To calculate the

sensitivity of the between-group distance, let $D_{orig} = \left| \sum_j E_{j1} \left(\bar{x}_{j(1)} - \bar{x}_{j(2)} \right) \right|$ and $D_{new} = \left| \sum_j E'_{j1} \left(\bar{x}_{j(1)} - \bar{x}_{j(2)} \right) \right|$. Define $\Gamma'_i = 10 \cdot \left(D_{new} - D_{orig} \right) / D_{orig}$

Criterion C: Correlation. For this criterion, we assume that the correlation of the original variable with the significant canonical axes produced by the MPGICA determines its importance. In criterion C, we set a significance limit of α. If the canonical axis has $P_{\chi^2}(i) < \alpha$, then it is included. The correlation of variable *k* is found for each of these significant axes and the maximum correlation is used. That is,

$$RS_{C,k} = corr\left(x_k, \mathbf{e}^{\mathbf{T}i}\mathbf{x}\right) = \left[\sum_{j=1}^{N} \left(x_{kj} - \bar{x}_k \right) \sum_{l=1}^{p} E_{il}^T \left(x_{lj} - \bar{x}_l \right) \right] \Bigg/ \sqrt{ \sum_{j=1}^{N} \left(x_{kj} - \bar{x}_k \right)^2 \sum_{m=1}^{N} \left[\sum_{l=1}^{p} E_{il}^T \left(x_{lm} - \bar{x}_l \right) \right]^2 }$$

.

BW ratio of Dudoit et al. (2002). Ramaswamy et al. (2003) used a univariate signal-to-noise ratio $S_x = \left(\mu_1 - \mu_2 \right) / \left(\sigma_1 + \sigma_2 \right)$ where the subscripts refer to group number, μ is the group mean, and σ is the standard deviation within the group. For cases in which the number of groups exceed two, we generalize the signal-to-noise ratio to the ratio of the between-group variance to the within group variance for each variable *i*. $S_i^2 = \left[\sum_{k=1}^{h} n_{ki} \left(x_{i(k)} - \bar{x}_i \right)^2 \right] \Big/ \sum_{j=1}^{N} \left(x_{ij} - \bar{x}_{i(k)} \right)^2 = (\mathbf{B})_{ii} / (\mathbf{W})_{ii}$. This is the BW ratio of Dudoit et al. (2002). Note that this criterion does not use any information from the MPGI canonical analysis.

Section IX. Inference Tests Used in the Generalized Inverse, Backward Elimination, and Forward Selection Canonical Analyses Algorithms

Overall group differences. We use the *F* approximation due to Rao (1965, p. 471) for the distribution of Wilks ratio to find the probability of error for rejection of the null hypothesis that all group means are identical. In the standard canonical analysis when variable selection is not an issue, this test gives an indication of the overall nature of the group separation. Wilks ratio is given by $\Lambda_l = \det|\mathbf{W}| / \det|\mathbf{W} + \mathbf{B}|$ for *l* variables. Denote the probability found from this test as $P_{FW}(l)$.

Significance of eigenvectors (*canonical axes*). The hypothesis tested is that the means of the groups all lie in an *r*-dimensional hyperplane. The test works by sorting the eigenvalues largest to smallest. The test of significance is made sequentially on all the eigenvalues remaining to be tested. So the first test is the probability of error in accepting all the *n* eigenvectors as a group. If that test is significant, then the test is made on eigenvalues 2 through *n*. If these cannot be rejected as a group, then the next test is made on eigenvalues 3 through *n*, and so on. At the *k*th test, we test for whether the

(*n*–*k*+1) remaining eigenvectors significantly discriminate using Bartlett's chi-square approximation. We denote the probability of χ^2 for eigenvalues j through *n* as $P_{\chi^2}(j)$. This test is in standard texts, e.g., Cooley and Lohnes (1971, p. 249) or Mardia et al. (1979, p. 343).

Significance of individual variables: coefficients of eigenvectors. We test for significance of each coefficient of the significant eigenvectors in the generalized inverse, backward elimination, and forward selection canonical analyses. Rao (1970) points out that individual coefficients of eigenvectors are not defined. Only ratios of coefficient of eigenvectors are defined. As a consequence, Rao asserts we can only test on the significance of a coefficient relative to other coefficients. So suppose that (*i*–1) variables out of $p \le (N-h)$ are included in a group used in a multivariate analysis. Rao (1970) originally developed this test for two groups. Hawkins (1976) and McHenry (1978) extended the test to multiple numbers of groups. Define Λ_{i-1} as Wilks ratio for the group of (*i*–1) variables. To test if a new particular variable is significant enough to be added to the group, denote Wilks ratio for the new set, which includes this new variable added to the previous (*i*–1) variables as Λ_i. Define $t_i = \Lambda_i / \Lambda_{i-1}$. Then the *F*-statistic for the significance of this particular *i*th variable is given by $F = \left[(1-t_i)/t_i\right](N-h-i+1)/(h-1)$ where $\nu_1 = h-1$ and $\nu_2 = N-h+i+1$ with probability denoted by $P_{Ft}(i)$. The implementations of the stepwise procedures calculate this statistic with and without the Bonferroni adjustment.

Bonferroni adjustment of $P_{Ft}(i)$. Bonferroni adjustments are known to be conservative. In fact they are sometimes referred to as a "protection rate" (e.g., Hays 1994, p. 451). Hays (1994 p. 451) suggests that in "current practice" they are often ignored for small numbers of tests. The most conservative approach, if Type I errors are of primary concern, is to use the Bonferroni adjustments always, no matter whether the forward selection or backward elimination is used or what the value of (*N*–*h*) is. The few cases analyzed to date suggest that it is more desirable to use the Bonferroni adjustment in the forward selection CA than in the backward elimination CA. Also, as the number (*N*–*h*) of variables in the "short list" increases, the desirability of using the Bonferroni adjustment also increases.

MPGI canonical analysis procedure. After reading in the data matrix **X**, we form the within-group data matrix $\mathbf{X_W}$. We perform an SVD on $\mathbf{X_W}$ using the routine DSVDC from the Slatec Library Version 4.1. We use DSVDC to calculate only the first *N* columns of **V** and the eigenvalue matrix $\Delta_{\mathbf{W}}^{1/2}$. The eigenvector $\mathbf{s}^1$ can be calculated directly from the group means as shown in Supplement Section VII. The vector $\mathbf{e}'$ is found from eq. 4 and normalized by eq. 3 to the vector $\mathbf{e}^1$. Each observation is then projected onto the canonical axis, using eq. 1. The chi-square tests are calculated using the Slatec Library version 4.1 routine DGAMIC. The probabilities for the *F*-distribution are calculated usting the Slatec Library version 4.1 routine DBETAI. We then calculate and display the four criteria for each of the original variables as described above. Correlations for criteria C are calculated using the Slatec Library Version 4.1 routine

DCOVAR. We then sort the variables according to the chosen criterion. We cut off this list at (N–h) variables.

Backward elimination algorithm. Either backward elimination or forward selection is then performed on this list of (N–h) variables. In the backward elimination procedure, we solve eq. 2 using RSG from the Slatec Library Version 4.1. At step i in the back elimination program, (N–h –i+1) variables are in the current list. We calculate the current value of Wilks ratio (or use the one calculated in the previous step for this list). We then temporarily remove each variable from the current list, calculate Wilks ratio, replace the variable and go to the next, until a Wilks ratio has been calculated with each variable removed. The variable associated with Wilks ratio closest in value to the current value of Wilks ratio is removed. The significance of this variable is calculated. This variable is the worst performing variable of the current set. The canonical equation is solved for this step, all statistical tests are calculated, and the positions of all observations in canonical space are displayed. Sensitivity analysis and correlations of all variables with the canonical axes are calculated. This completes one step. The algorithm iteratively repeats this step (N–h–1) times. Then a canonical analysis is performed on the one remaining variable for a final step.

Forward selection. The forward selection algorithm is nearly the reverse of the backward elimination algorithm. Instead of eliminating a variable at each step, we begin with no variables in the current set and add a variable at each step. We add the variable that produces the most improvement in Wilks ratio. Significance tests and canonical analyses are conducted at each step. The first several steps are performed by exhaustive search rather than by strict addition of one new variable. The user determines the number of steps performed by exhaustive search up to a maximum number of ten steps.

REFERENCES

Ambroise, C., G.J. McLachlan. 2002. Selection bias in gene expression on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences 99:6562-6566.

Campbell, S.L., C.D. Meyer, Jr. 1979. Generalized inverses of linear transformations. Pitman: London.

Cooley, W.W., P.R. Lohnes. 1971. Multivariate data analysis. John Wiley & Sons.

Dudoit, S., J. Fridlyand, T.P. Speed. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Statistic. Assoc. 97:77-87

Furey, T.S., N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16:906-914.

Golub, R.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeck, J.P. Mesirov, H. coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286:531-537.

Hawkins, D.M. 1976. The subset problem in multivariate analysis of variance. J. Royal Statist. Soc. B 38:132-139.

Hays, W.L. 1994. Statistics. 5th ed. Wadsworth Thomson Learning.

Kari, L., A. Loboda, M. Nebozhyn, A.H. Rook, E. C. Vonderheid, C. Nichols, D. Virok, C. Chang, W.H. Horng, J. Johnston, M. Wysocka, M. K. Showe, L.C. Showe. 2003. Classification and prediction of survival in patients with the leukemic phase of cutaneous T cell lymphoma. Journal of Experimental Medicine 197:1477-1488.

Kercher, J.R., R.G. Langlois, B.A. Sokhansanj, C.F. Melius, J.N. Quong, F.P. Milanovich, A.A. Quong. 2004. Variable selection in canonical analysis of gene- and protein-expression data: the special case of two groups. (Submitted for publication).

Kozak, K.R., M.W. Amineus, S.M. Pusey, F. Su, M.N. Luong, S.A. Luong, S.T. Reddy, R. Farias-Eisner. 2003. Identification of biomarkers for ovarian cancer using storn anion-exchange ProteinChips: potential use in diagnosis and prognosis. Proceedings of the National Academy of Science 100:12343-12348.

Krzanowski, W.J. 2000. Principles of multivariate analysis. Clarendon Press, Oxford.

Langlois, R.G, J.E. Trebes, E.A. Dalmasso, Y. Ying, R.W. Davies, M.P. Curzi, B.W. Colston Jr., K.W. Turteltaub, J. Perkins, B.A. Chromy, M.W. Choi, G.A. Murphy, J.P. Fitch, and S.L. McCutchen-Maloney. 2004. Serum protein profile alterations in hemodialysis patients. American Journal of Nephrology (in Press)

Mardia, K.V., J.T. Kent, J.M. Bibby. 1979. Multivariate analysis. Academic Press.

McHenry, C.E. 1978. Computation of a best subset in multivariate analysis. Appl. Statist. 27:291-296.

Meyer, C.D. Jr. 1973. Generalized inversion of modified matrices. SIAM J. Appl. Math. 24:315-323.

Moler, E.J., M.L. Chow, I.S. Mian. 2000. Analysis of molecualr profile data using generative and discriminative methods. Physiological genomics 4:109-126.

Nguyen, D.V., D.M. Rocke. 2002. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics 18:1216-1226.

Radmacher, M.D., L.M. McShane, R. Simon. 2002. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 9:505-511

Ramaswamy, S., K.N. Ross, E.S. Lander, T.R. Golub. 2003. A molecular signature of metastasis in primary solid tumors. Nature Genetics 33:49-54.

Rao, C.R. 1965. Linear statistical inference and its applications. John Wiley, New York.

Rao, C.R. 1970. Inference on discriminant function coefficients. P. 587-602. *In* Essays on Probability and Statistics (R.C. Bose, I.M. Chakravarti, P.C. Mahalanobis, C.R. Rao, K.J.C. Smith, eds) University of North Carolina Press:Chapel Hill, NC.

Schott, J.R. 1997. Matrix analysis for statistics. John Wiley & Sons : New York.

Seal, H. 1964. Multivariate statistical analysis for biologists. Wiley.

Srivastava, M.S. 2002.Methods of multivariate statistics. Wiley.

Wright, G., B. Tan, A. Rosenwald, E.H. Hurt, A. Wiestner, L.M. Staudt. 2003. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proceedings of the National Academy of Sciences 100:9991-9996.

Xiong M.M., L. Jin, W.J. Li, E. Boerwinkle. 2000. Computational methods for gene expression-based tumor classification. Biotechniques 29: 1264-1270.

Xiong, M. W. Li, J. Zhao, L. Jin, E. Boerwinkle. 2001. Featgure (gene) selection in gene expression-based tumor classification. Molecular Genetics Metabolism 73:239-247.

Zhang, H., C.-Y.Yu, B. Singer, M. Xiong. 2001. Recurseive partititioning for tumor classification with gene expression microarray data. Proc. National Academy of Science 98:6730-6735.

Supplement to Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The Special Case of Two Groups
PART B: Results

J. R.Kercher[1], R.G. Langlois[2], B.A. Sokhansanj[3],
C.F. Melius[4], J.N. Quong[4], F.P. Milanovich[5], B.W. Colston, Jr.[6],
K.W. Turteltaub[2], A.A. Quong[3]

[1]Environmental Sciences Division, L-235,
[2]Biodefense Division, [3]Chemical Biology and Nuclear Science Division,
4Chemistry and Chemical Engineering Division, [5]R Division
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, California 94551

{NOTE TO THE EDITOR: If the main paper is accepted, we understand this supplemental report will be placed online.)

Section X. Data handling and the effect of scale change on contributions to intergroup distance from Dialysis study.

Data handling. The data are measurements of 165 proteins in blood serum of eight dialysis patients and nineteen control subjects. See Langlois et al. (2004) for descriptions of sample preparation and handling. Rules-Based Medicine, Inc.™ (RBM) measured the 161 protein concentrations using reagent-coated fluorescent micro-sphere technology. Analyte levels below the limit of detection of the analysis were reported as "LOW". These LOW values were set to zero for the present analysis. We eliminated four variables from consideration that had either no or one non-zero value, leaving 161 proteins or protein ligands for canonical analysis.

In Fig. S.1 we show the effect of scale change on the contribution that each variable makes to the distance between the means of the two groups in canonical space. These figures are in addition to the graphs for ferritin and stem cell factor shown in Kercher et al. (2004).

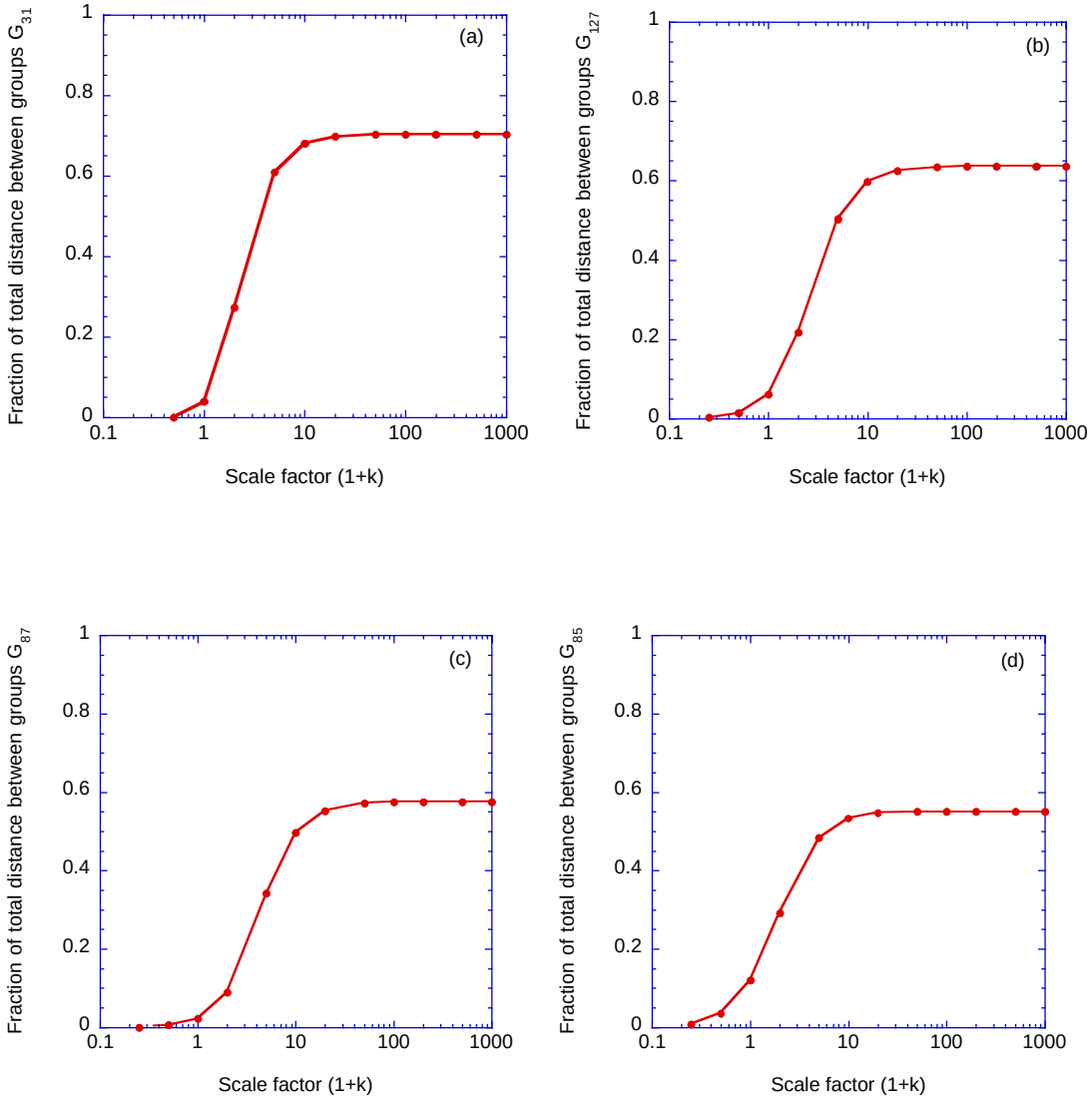


Fig. S.1 See caption next page.

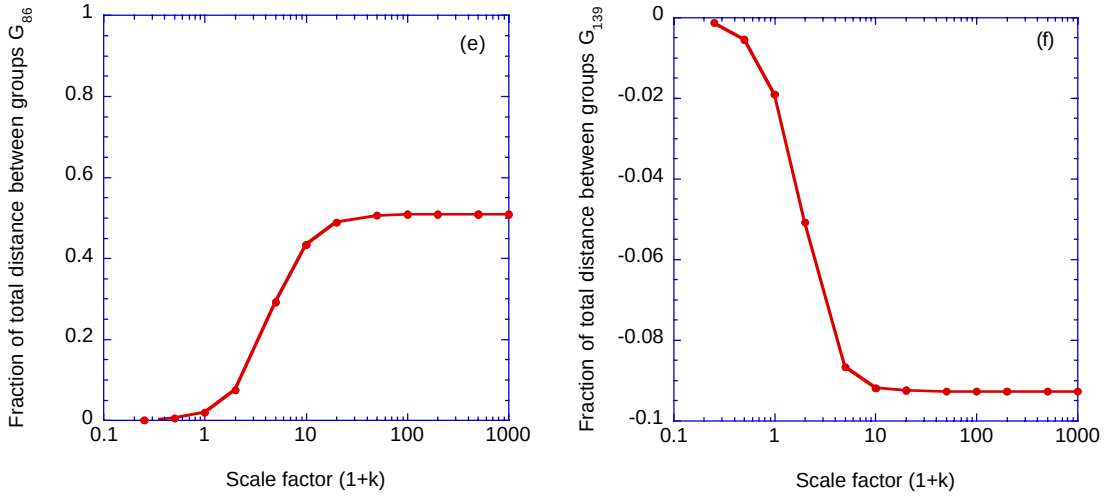


Fig. S.1. (Continued). These are additional graphs of the effect of changes in the scale factor (1+*k*) on the contribution of each variable to the distance between the two groups in canonical space for the RBM data of the dialysis experiment. Graphs for the two variables with the largest asymptote fraction, ferritin and stem cell factor, are shown in Kercher et al. (2004). Figs. S.1a, S1.b, S1.c, S1.d, and S1.e are for the five next largest asymptotes. (a) Variable 31 in the data set used in the canonical analyses is interleukin-16. (b) Variable 127 is Cytochrome P450. (c) Variable 87 is Hepatitis E Virus (orf2 6KD). (d) Variable 85 is Hepatitis A. (e) Variable 86 is Hepatitis E Virus (orf2 3KD). (f) is the graph of the variable with the largest negative asymptote. Note the vertical scale is expanded by a factor of 10 in (f) compared to the other figures. (f) Variable 139 is HSP 90 α.

Table S.1. The maximum fraction of the distance between the two groups, i.e., dialysis patients and control group, (max G_i) due to the variable for large scale factor k. The variable number in columns one and four refer to its location in the original data set.

No.	Variable	Max G_i	No.	Variable	Max G_i
18	Ferritin	0.821	7	Brain-Derived_Neurotrophic _Factor	0.394
57	Stem_Cell_Factor	0.731	109	Parainfluenza_1	0.371
32	IL-16	0.704	95	HIV-1_gp41	0.369
131	Cytochrome_P450_	0.636	137	Histone_H3	0.363
91	Hepatitis_E_Virus_(orf2_6KD)	0.575	2	Alpha-Fetoprotein	0.354
89	Hepatitis_A_	0.551	115	Rubeola	0.333
90	Hepatitis_E_Virus_(orf2_3KD)	0.510	112	Polio_Virus	0.330
60	Tissue_Factor	0.486	101	Influenza_A	0.284
52	Myoglobin	0.405	145	Insulin	0.282
92	Hepatitis_E_Virus_(orf3_3KD)	0.396	61	TIMP-1	0.281

Section XI. Selecting canonical variates in RBM Dialysis data.

In Fig. S.2 we show the graphs of the *P*-value for the Rao-Hawkins-McHenry statistic on the significance of added or eliminated variable in the stepwise procedures.

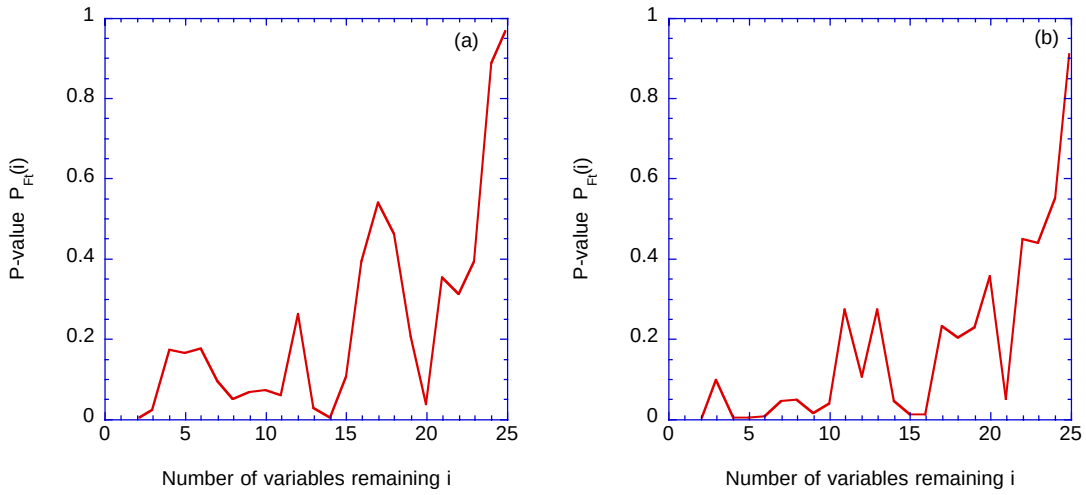


Fig. S.2. Graphs of the *P*-value for the test of the null hypothesis that the variable added or eliminated does not contribute significantly to the discrimination between the groups. (a) MPGICA-B-BECA (Criterion B: Absolute sensitivity). (b) MPGICA-C-BECA (Criterion C: Correlation). Data for RBM dialysis study.

Comment. At each step eliminating a variable, the ratio $\det|\mathbf{W}|/\det|\mathbf{W}+\mathbf{B}|$ degrades monotonically (gets larger). However, how much it gets larger can fluctuate, depending on the linear combinations of the variables in the old current list and the variables in the new current list.

Table S.2. Variables at cutoff for the two different stepwise procedures and four different criteria. Coefficients of the transforming eigenvector $\mathbf{e}^1$ for these variables are in column two and four. These are the results of stepwise canonical analyses on 161 proteins measured for eight dialysis patients and nineteen control subjects.

Variable	Coefficient of $\mathbf{e}^1$	Variable	Coefficient of $\mathbf{e}^1$
MPGICA-A-BECA: Criterion A: Sensitivity o Wilks ratio		**MPGICA-B-BECA: Criterion B Absolute Sensitivity of Wilks ratio**	
Ferritin	1.874	Ferritin	1.272
IL-16	0.913	Hepatitis_E_Virus_(orf2_6KD)	1.044
Hepatitis_E_Virus_(orf2_6KD)	0.831	Stem_Cell_Factor	0.459
Alpha-Fetoprotein_	–0.637		
MPGICA-C-BECA: Criterion C Correlation with canonical axis		**BW-BECA Signal-to-noise ratio pre-filter (conventional)**	
Ferritin	1.385	Ferritin	1.114
Hepatitis_E_Virus_(orf2_6KD)	1.142	Parainfluenza_1	0.701

Table S.3. Variables at cutoff for the two different stepwise procedures and four different criteria for RBM protein data for dialysis study. Sensitivities for these variables are in column two and four. Sensitivities are the relative change in the Wilks ratio statistic per relative change in the coefficient in $\mathbf{e}^1$ for the variable. Formal expressions for the sensitivities are given in the Supplement in Section VIII.

Variable	Sens-itivity	Variable	Sens-itivity
MPGICA-A-BECA:		**MPGICA-B-BECA**	
Ferritin	1.65	Ferritin	1.37
IL-16	0.348	Hepatitis_E_Virus_(orf2_6KD)	0.333
Hepatitis_E_Virus_(orf2_6KD)	0.215	Stem_Cell_Factor	0.327
α-Fetoprotein_	-0.161		
MPGICA-C-BECA		**BW-BECA (conventional)**	
Ferritin	1.66	Ferritin	1.79
Hepatitis_E_Virus_(orf2_6KD)	0.383	Parainfluenza_1	0.245

Table S.4. Correlations of the fifteen variables most positively correlated with the canonical axes for MPGICA-X-BECA and BW-BECA (conventional) in the dialysis study using RBM protein data. These variables increase in the direction of the dialysis patients.

Protein, Ligand, or Antibody, etc	Correl-ation	Protein, Ligand, or Antibody, etc	Correl-ation
MPGICA-A-BECA		**MPGICA-B-BECA**	
Ferritin	0.986	Ferritin	0.988
Stem_Cell_Factor	0.965	Stem_Cell_Factor	0.969
IL16	0.926	Cytochrome_P450_	0.919
Cytochrome_P450_	0.913	IL-16	0.913
Tissue_Factor	0.897	Tissue_Factor	0.902
Myoglobin	0.892	Myoglobin	0.894
Histone_H3	0.871	Histone_H3	0.880
Hepatitis_A_	0.862	Hepatitis_A_	0.872
Hepatitis_E_Virus_(orf2_6KD)	0.855	α-Fetoprotein_	0.863
α-Fetoprotein_	0.852	Hepatitis_E_Virus_(orf2_6KD)	0.856
Hepatitis_E_Virus__(orf2_3KD)	0.844	HIV-1_gp41	0.847
HIV-1_gp41	0.832	Hepatitis_E_Virus__(orf2_3KD)	0.841
HSC_70	0.830	HSC_70	0.841
Sc-l70	0.826	Scl-70	0.838
Rubeola	0.818	Rubeola	0.823
MPGICA-C-BECA		**BW-BECA (conventional)**	
Ferritin	0.989	Ferritin	0.994
Stem_Cell_Factor	0.956	Stem_Cell_Factor	0.948
Cytochrome_P450_	0.922	Tissue_Factor	0.911
IL-16	0.911	Cytochrome_P450_	0.909
Tissue_Factor	0.908	Myoglobin	0.896
Myoglobin	0.892	IL-16	0.885
Histone_H3	0.878	α-Fetoprotein_	0.869
α-Fetoprotein_	0.872	Histone_H3	0.863
Hepatitis_A_	0.871	HIV-1_gp41	0.838
Hepatitis_E_Virus_(orf2_6KD)	0.855	HSC_70	0.836
HIV-1_gp41	0.849	Hepatitis_A_	0.831
Hepatitis_E_Virus__(orf2_3KD)	0.846	Hepatitis_E_Virus_(orf2_6KD)	0.815
HSC_70	0.842	HSP_32_(HO)	0.815
Scl-70	0.832	Rubeola	0.803
Rubeola	0.810	Influenza_A	0.802

We have divided each of the four lists in Table S.4 into six sections: variables 1 and 2; 3 and 4; 5, 6, and 7; 8, 9, and 10; 11 and 12; and 13, 14, and 15. For the three new two-stage MPGICA-X-BECA's in Table S.4, the variables are identical in each section; in some sections the rankings of the variables are identical. However, except for the first

section, the memberships are not the same for the conventional method as for the new method in any of the sections. In fact, two variables (Hepatitis E orf2,3KD and ScI-70) are dropped from the list of the new method and replaced by two others (HSP-32 and Influenza A) in the list for the conventional method.

Table S.5. Correlations of the ten variables most negatively correlated with the canonical axis for both stepwise procedures for all four criteria. These variables increase in the direction of the control subjects. These results are for RBM data for eight dialysis patients and 19 control subjects tested for 161 proteins.

Protein, Ligand, or Antibody, etc	Correl-ation	Protein, Ligand, or Antibody, etc	Correl-ation
MPGICA-A-BECA		MPGICA-B-BECA	
Brain-Derived_Neurotrophic_Factor	–0.814	Brain-Derived_Neurotrophic_Factor	–0.816
Thrombopoietin	–0.557	RANTES	–0.545
RANTES	–0.551	Thrombopoietin	–0.544
von_Willebrand_Factor	–0.496	von_Willebrand_Factor	–0.499
ENA-78_	–0.465	IL-7	–0.447
IL-7	–0.464	ENA-78_	–0.442
MMP-9_	–0.396	IL-15	–0.377
IL-15	–0.377	MMP-9_	–0.376
GM-CSF_	–0.362	GM-CSF_	–0.365
Apolipoprotein_CIII_	–0.361	Apolipoprotein_CIII_	–0.355
MPGICA-C-BECA		BW-BECA	
Brain-Derived_Neurotrophic_Factor	-0.812	Brain-Derived_Neurotrophic_Factor	-0.798
RANTES	-0.562	Thrombopoietin	-0.551
Thrombopoietin	-0.549	RANTES	-0.546
von_Willebrand_Factor	-0.505	von_Willebrand_Factor	-0.499
ENA-78_	-0.479	ENA-78_	-0.471
IL-7	-0.471	IL-7	-0.465
MMP-9_	-0.388	IL-15	-0.379
IL-15	-0.376	MMP-9_	-0.374
GM-CSF_	-0.361	Apolipoprotein_CIII_	-0.358
Apolipoprotein_CIII_	-0.351	GM-CSF_	-0.358

Table S.6. Spearman rank correlations of the list of correlations of the variables with the canonical axis for the new two-stage method using three different criteria for ranking and the conventional BW-BECA with the BW ratio as a pre-filter. These results are for the dialysis data.

Procedure/Criterion	MPGICA-A-BECA	MPGICA-B-BECA	MPGICA-C-BECA
MPGICA-B-BECA	0.999		
MPGICA-C-BECA	0.999	0.999	
BW-BECA	0.991	0.993	0.993

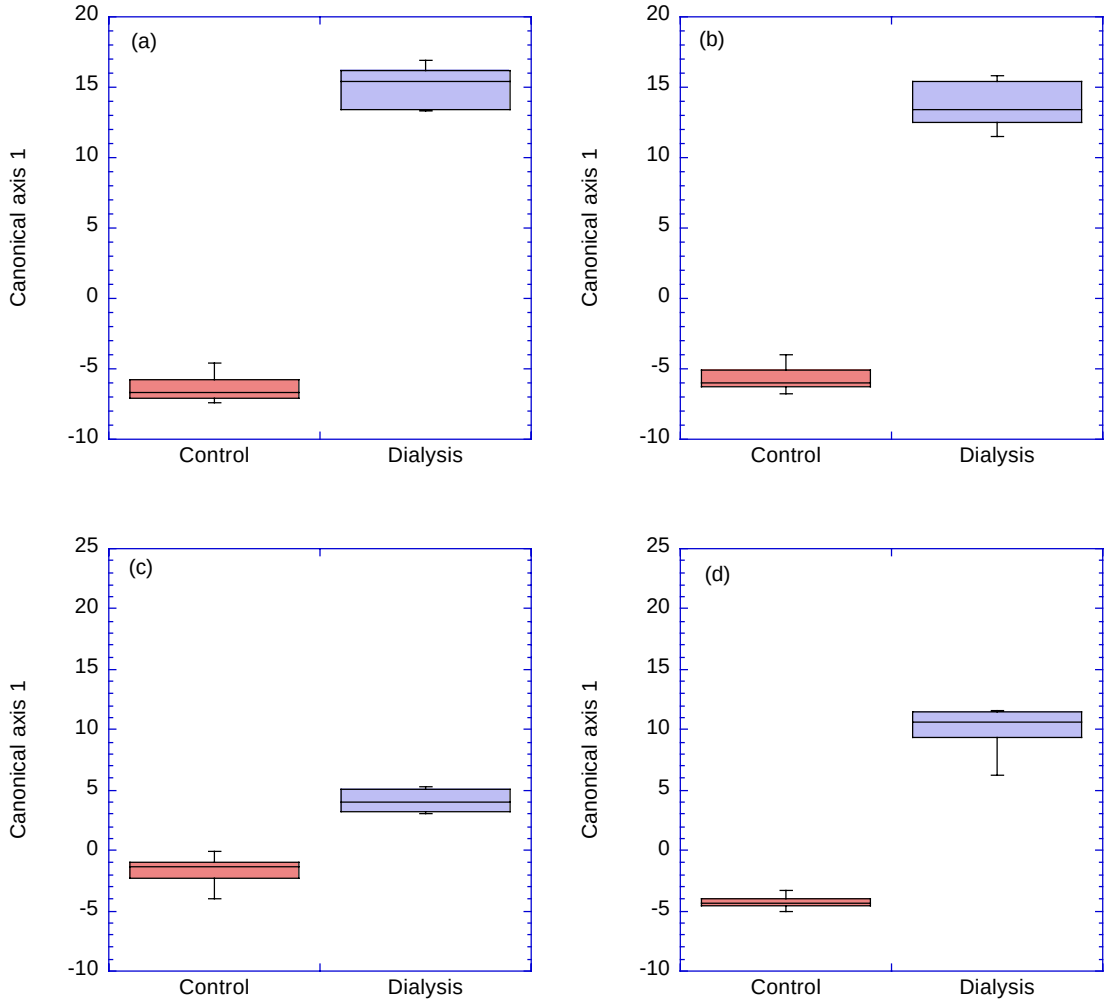


Fig S.3. Boxplots of RBM dialysis data (161 proteins) canonical analyses. (a) MPGICA-B-BECA. Criterion B: Absolute sensitivity. Three proteins. (b) MPGICA-C-BECA. Criterion C: Correlation. Two proteins. (c) MPGICA only. No followup backward elimination. (d) BW-BECA (conventional backward elimination with BW (Dudoit et al. 2002) ratio pre-filter. Two proteins. See Table S.3 for the lists of proteins.

Supplement Section XII. Forward selection on the full set of variables with no MPGI filter.

Consider the application of a pure forward selection procedure to the data for 161 variables in the dialysis study. See Fig. S.4 for $P_{Ft}(i)$. At each step, we allow the procedure to examine each of the variables remaining from the original set of 161, which are not already included in the current set of variables. We find that cutoff occurs at 23 variables out of a maximum possible number of variables that can be used of $(N–h)=25$. The resulting separation of the groups is an unrealistically high overfit. See boxplots in Fig. S.5. One solution to avoid this "capitalization on chance" is to use a Bonferroni adjustment to the RHM test (Hawkins 1976) (Fig. S.6). This adjustment reduces the number of variables at cutoff to four. See Fig. S.7 for new boxplot results. We suggest that the Bonferroni adjustment has substantially reduced the overfit, but has not eliminated it. Because of the many variables to choose from, the Bonferroni-adjusted FSCA can still exploit variation within the groups to construct linear combinations of variables that reduce group width, leading to an apparent increase in group separation. As shown in Table 2 in Kercher et al. (2004), the four variables used by the MPGICA-A-BECA are all highly correlated with the canonical axis, and all have signal-to-noise ratios greater than 1.54. However, three of the four variables for the Bonferroni-adjusted FSCA are weakly correlated with the axis and these three each have signal-to-noise ratios less than 0.57. These results suggest that in this instance more exploitation of random fluctuations is occurring in the Bonferroni-adjusted FSCA than in the MPGICA-A-BECA.

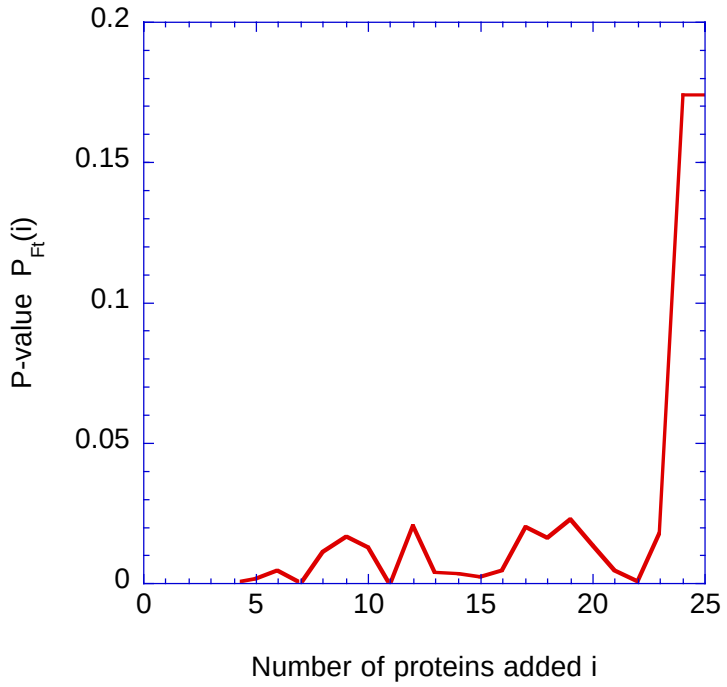


Fig. S.4. Graph of the *P*-value implied by the Rao-Hawkins-McHenry statistic for forward selection using all 161 RBM variables for the dialysis data of eight dialysis patients and 19 control subjects. This test was originally devised for the number of variables being less than the number of observations less the number of groups, i.e., $p \leq (N-h)$ for nominal test. We ignore this restriction and proceed formally. Note cutoff is at 23 variables.

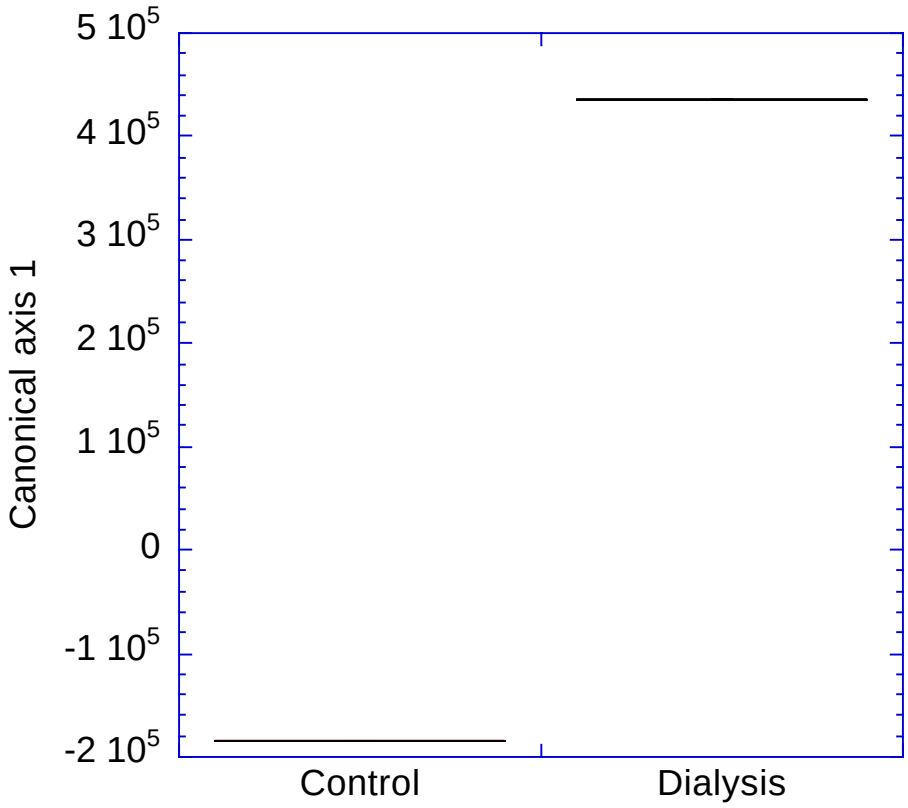


Fig. S.5. Boxplots of the two groups (eight dialysis patients, 19 control subjects) using 161 proteins measured by RBM in forward selection canonical analysis. We did not use the MPGI canonical analysis as a filter. Instead we relied solely on Rao-Hawkins-McHenry test to find cutoff variable. The resulting unrealistic overfit of the groups is due to "capitalization on chance" by the forward selection algorithm, which selects variables whose linear combination will reduce the size of the clusters in canonical space.

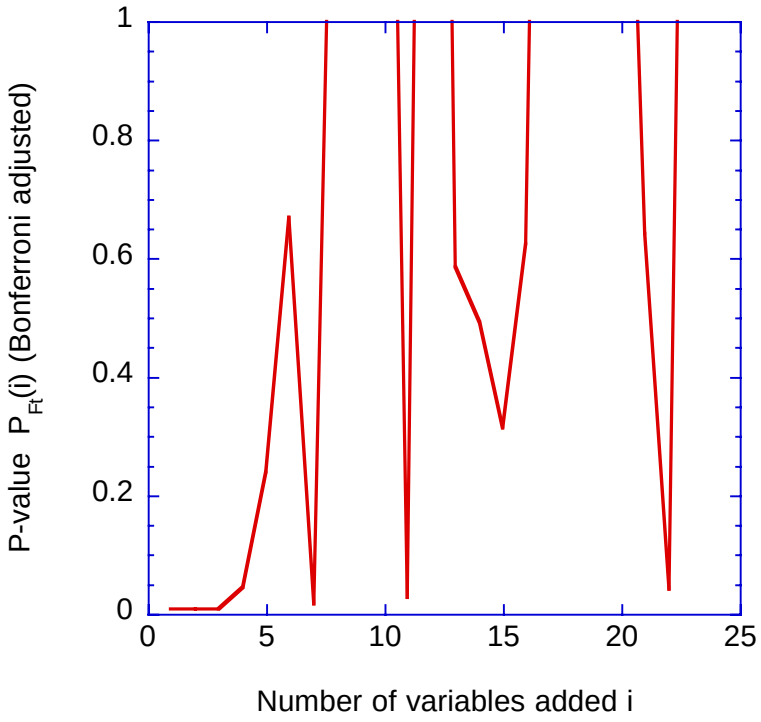


Fig. S.6. Bonferonni-adjusted *P*-value for forward selection using all 161 variables with no MPGI CA filter. Note that the cutoff variable is now four rather than the 23 as shown in Fig. S.4. Analysis for dialysis study of eight dialysis patients and nineteen control subjects.

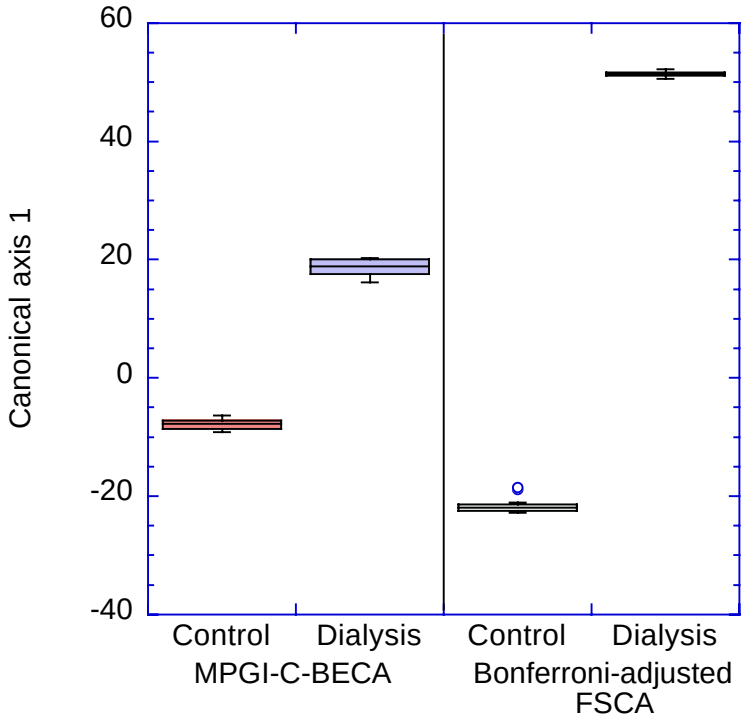


Fig. S.7. Comparison of group separation from MPGICA-A-BECA to the group separation from a pure forward selection CA using Bonferroni adjustment of the Rao-Hawkins-McHenry test. CA's use RBM protein data from dialysis study. In canonical units, the separation of the two groups for MPGICA-A-BECA and Bonferroni-adjusted FSCA-only is 27 and 73, respectively.

Section XIII. Separation of primary and metastases tumors in canonical space.

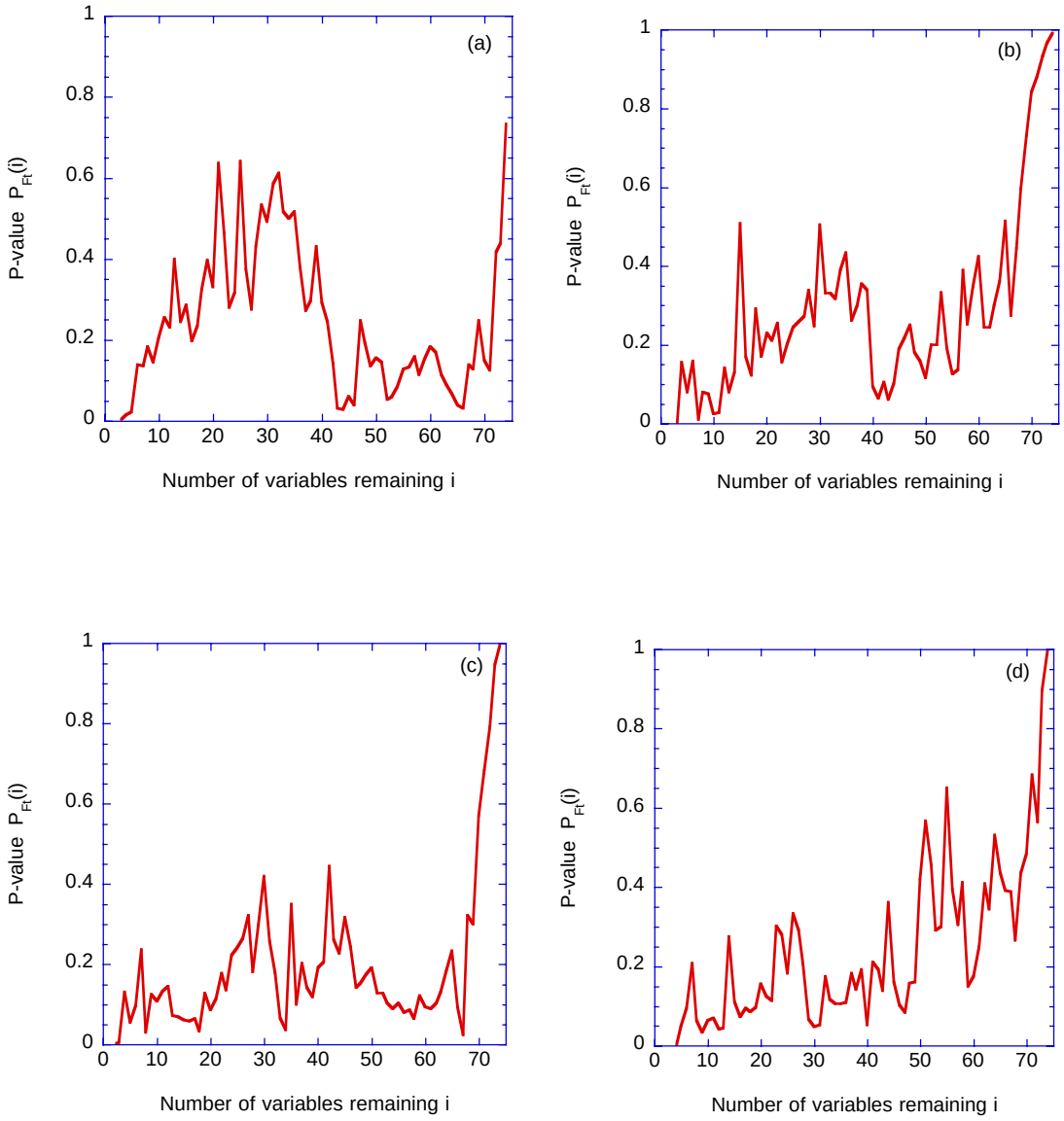


Fig. S.8. Plots of the *P*-value $P_{Ft}(i)$ implied by the Rao-Hawkins-McHenry statistic for rejecting the null hypothesis that the eliminated variable does not significantly improve discrimination between the two groups: primary tumors and metastases. These graphs are from backward elimination canonical analysis following MPGICA where the variables were ranked according to (a) criterion A: Sensitivity of Wilks ratio, (b) criterion B: Absolute sensitivity of Wilks ratio, and (c) criterion C: Correlation of the variable with the canonical axis. (d) Conventional BECA with univariate BW ratio pre-filter. We select canonical analyses at 5, 11, 8, and 12 variables remaining for further examination. Analysis of 128-gene data set derived from Dataset A from Ramaswamy et al. (2003).

Table S.7. Transformation vectors $\mathbf{e}^1$ ($\times 10^{-2}$) for Bonferroni-adjusted cutoff **B**, unadjusted cutoff **C**, and low-end variables **L** for MPGICA-A,B,C-BECA and BW-BECA (conventional). Data consists of 128 genes from Dataset A of Ramaswamy et al. (2003) for 64 primary tumors and 12 metastases.

Hu6800/HU35KsubA Accession	GenBank Accession	**B** $\mathbf{e}^1$ terms	**C** $\mathbf{e}^1$ terms	**L** $\mathbf{e}^1$ terms
MPGICA-A-BECA : Sensitivity of Wilks ratio				
RC_AA460436_at	RC_AA460436_at	1.345	0.963	N.A.
U65410_at	U65410_at	4.220	4.599	N.A.
Z74615_at	Z74615_at	0.030	0.020	N.A.
K03515_at	K03515_at		0.052	N.A.
AA412620_s_at	AA412620_s_at		0.371	N.A.
MPGICA-B-BECA: Absolute sensitivity of Wilks ratio				
X82494_at	X82494	0.756	0.756	1.209
RC_AA195031_at	AA195031	0.198	0.198	0.222
Z14244_at	Z14244	0.162	0.162	0.169
RC_AA608850_at	AA608850			0.159
RC_AA449951_at	AA449951			–0.318
RC_AA400410_at	AA400410			0.347
RC_AA100089_at	AA100089			–0.226
Z74616_s_at	Z74616			–0.046
RC_AA037386_s_at	AA037386			–0.177
RC_AA412059_at	AA412059			0.364
X85372_at	X85372			0.393
MPGICA-C-BECA: Correlation with canonical axis				
X82494_at	X82494	0.917	0.802	0.708
RC_AA460436_at	AA460436	1.515	1.481	1.475
U65410_at	U65410		2.750	5.652
RC_AA449951_at	AA449951			-0.304
L37747_s_at	L37747			1.174
RC_AA428024_at	AA428024			0.187
RC_AA037386_s_at	AA037386			-0.144
AA486831_s_at	AA486831			0.126
BW-BECA (conventional)				
S80437_s_at	S80437	0.075	0.067	0.103
U65410_at	U65410	4.902	4.621	8.642
K03515_at	K03515	0.081	0.079	0.078
D89377_i_at	D89377		0.749	0.605
RC_AA009596_at	AA009596			0.431
AA093131_at	AA093131			–0.237
AA486831_s_at	AA486831			0.200
RC_AA449951_at	AA449951			–0.408
L37747_s_at	L37747			1.521
J02783_at	J02783			–0.028
AA096094_s_at	AA096094			0.047
RC_AA037386_s_at	AA037386			–0.148

Table S.8. The fifteen genes with highest positive correlations and the ten genes with the highest negative correlations with the canonical axis for the 128-gene primary/metastases data set of Ramaswamy et al. 2003. J03464 increases in the direction of metastases. The correlation values were averaged over all criteria. These correlations were averaged over the five-gene, eleven-gene, and eight-gene CA's for criterion A, B, and C, respectively, listed in Table S.7.

GenBank ID	Positive Correlation	GenBank ID	Negative Correlation
J03464	0.632		
AA400410	0.626	HG4660-HT5073	-0.263
AA460436	0.624	M83664	-0.268
X82494	0.608	M26061	-0.271
AA195031	0.592	U45974	-0.272
AA252812	0.589	U45448	-0.285
AA025213	0.568	M27830	-0.287
AA428024	0.563	S67156	-0.296
U75285	0.563	S72043	-0.298
AA236972	0.551	D43968	-0.312
Z74616	0.548	X66141	-0.334
AA449951	0.543		
HG4264-HT4534	0.541		
X85372	0.541		
AA609674	0.540		

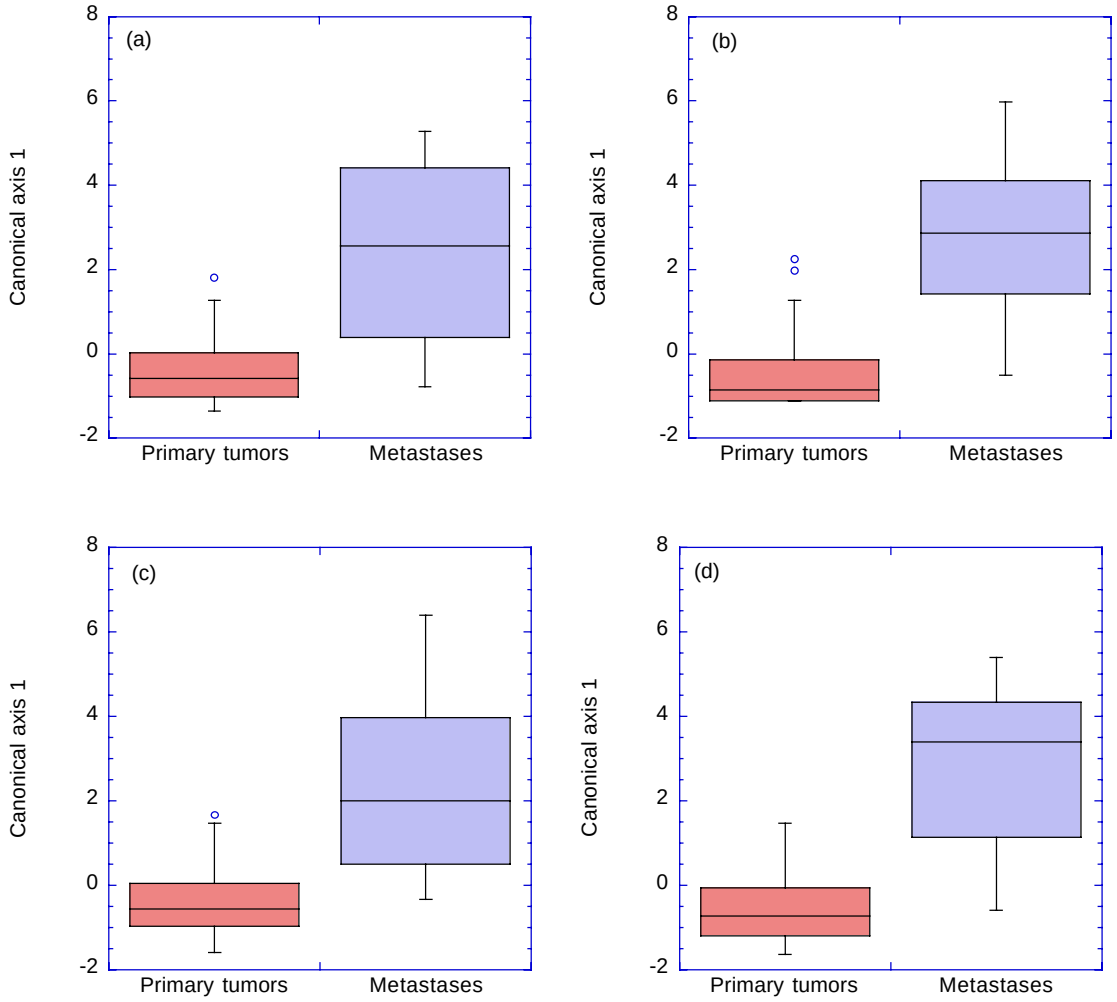


Fig. S.9. Boxplots of metastases and primary tumors following backward elimination. (a) MPGICA-A-BECA, Bonferroni-adjusted cutoff at three genes. (b) MPGICA-C-BECA, Bonferroni-adjusted cutoff at two genes. (c) BW-BECA, Bonferroni-adjusted cutoff at three genes. (d) MPGICA-A-BECA, unadjusted cutoff at five genes. Data from Ramaswamy et al. (2003) for 128 genes from their dataset A for 64 primary tumors and 12 metastases.

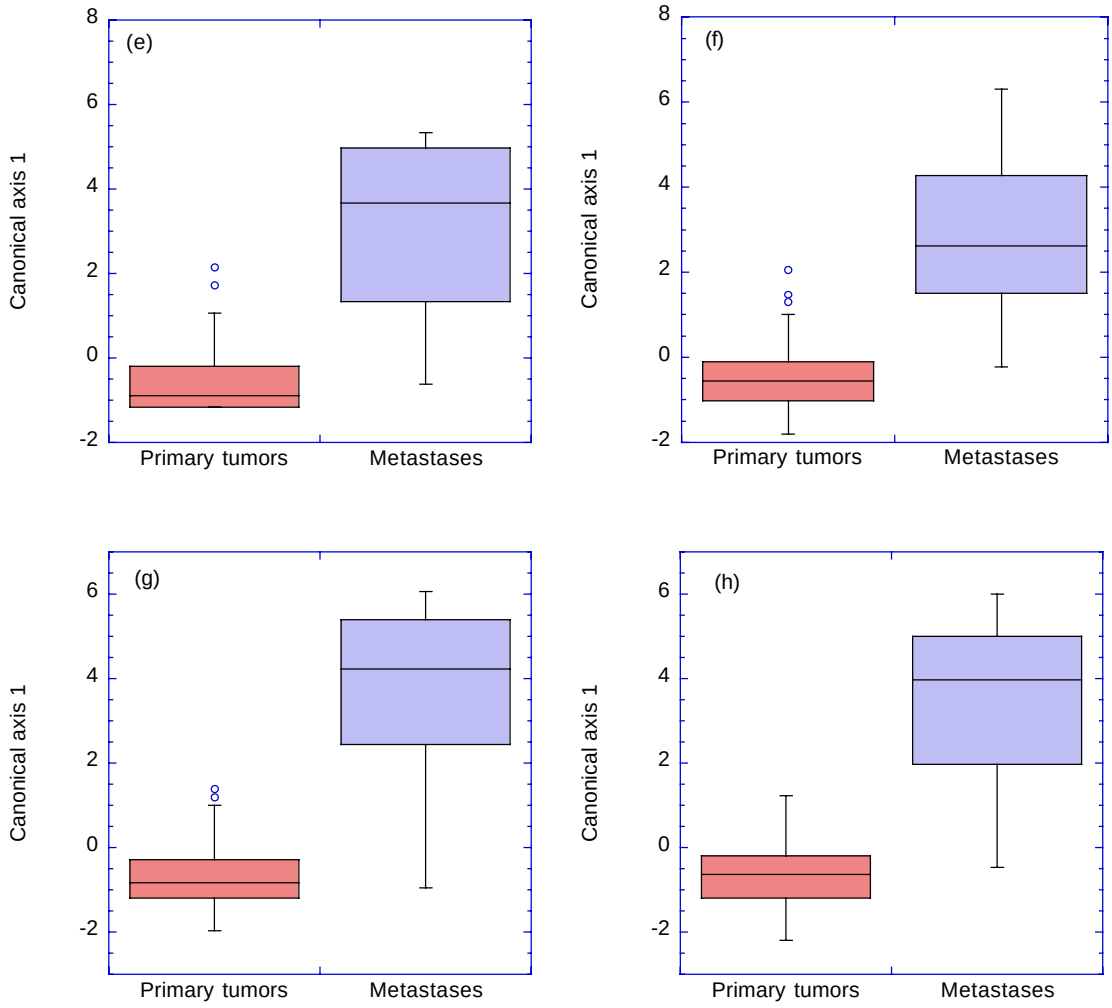


Fig. S.9 (continued). (e) MPGICA-C-BECA, unadjusted cutoff at three genes. (f) BW-BECA, unadjusted cutoff at four genes. (g) MPGICA-C-BECA, eight genes (low-end). (h) BW-BECA, twelve genes (low-end).

Section XIV. Lung data. Signatures for groups classified by clustering

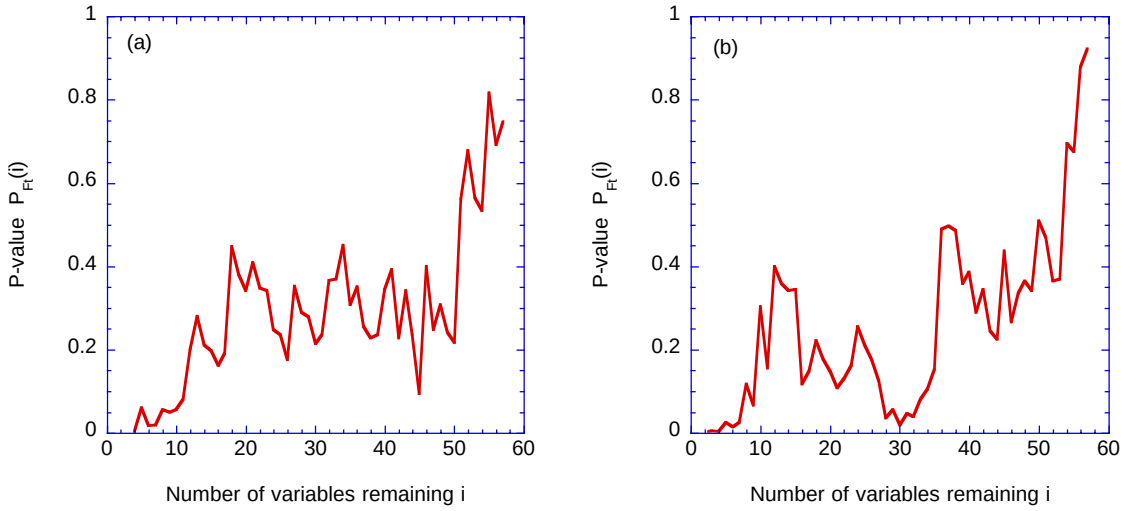


Fig. S.10. Significance of Rao-Hawkins-McHenry test ($P_{Ft}(i)$ plotted against i. We plot the P-value if we reject the null hypothesis that the coefficient of the eliminated variable is zero. These graphs are for the 169-gene, two-group case of primary lung tumors classified as either Cluster 0 or Cluster 1. These classifications result from clustering procedures. Data for canonical analyses are from Ramaswamy et al. (2003) Dataset B. (a) MPGICA-A-BECA. (b) BW-BECA.

Table S.9. Gene lists for canonical analyses, which separated the lung primary data into the two clusters specified by the 17-gene signature of Ramaswamy et al. 2003. The nine-gene list on the left is from MPGICA-A-BECA and the eleven-gene list on the right is from BW-BECA (conventional) with significance of eliminated variable at 0.05 and 0.15, respectively.

MPGICA-A-BECA			BW-BECA		
Affymetrix U95A Accession ID	GenBank ID	Coefficients of $\mathbf{e}^1$	Affymetrix U95A Accession ID	GenBank ID	Coeffiecients of $\mathbf{e}^1$
40412_at	AA430032	0.011362	41403_at	X85372	0.0132239
32229_at	AA608850	0.017517	32229_at	AA608850	0.0186209
38216_at	L40411	–0.046079	35138_at	AA279205	–0.0387291
39122_at	K03515	0.001449	38841_at	AA609674	0.0165823
38429_at	S80437	–0.001142	40825_at	AA448655	–0.0078906
32847_at	U48959	–0.007683	38654_at	AA412059	0.0229312
38095_i_at	M83664	–0.000688	40872_at	AC002115	0.0010463
691_g_at	J02783	–0.001306	AFFX_HUMRGE_M10098_M_at	M10098	0.0232126
35710_s_at	U95006	0.010342	41349_at	L43964	–0.0203426
			32825_at	Y10807	–0.0040433
			33433_at	AA037386	0.0069268

REFERENCES

Dudoit, S., J. Fridlyand, T.P. Speed. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Statistic. Assoc. 97:77-87.

Kercher, J.R., R.G. Langlois, B.A. Sokhansanj, C.F. Melius, J.N. Quong, F.P. Milanovich, A.A. Quong. 2004. Variable selection in canonical analysis of gene- and protein-expression data: the special case of two groups. (Submitted for publication).

Langlois, R.G, J.E. Trebes, E.A. Dalmasso, Y. Ying, R.W. Davies, M.P. Curzi, B.W. Colston Jr., K.W. Turteltaub, J. Perkins, B.A. Chromy, M.W. Choi, G.A. Murphy, J.P. Fitch, and S.L. McCutchen-Maloney. 2004. Serum protein profile alterations in hemodialysis patients. American Journal of Nephrology (in Press)

Ramaswamy, S., K.N. Ross, E.S. Lander, T.R. Golub. 2003. A molecular signature of metastasis in primary solid tumors. Nature Genetics 33:49-54.

UCRL-JRNL-205177

Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The General Case for Multiple Groups

J.R. Kercher , J.N. Quong, K.J. Wu, A.A. Quong

Lawrence Livermore National Laboratory, Livermore, California 94551

This article was prepared for journal submission.

July 2004

DISCLAIMER

Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The General Case for Multiple Groups

J. R.Kercher[1], J. N. Quong[1,2], K. J. Wu[1], A. A. Quong[1,2]

[1]Lawrence Livermore National Laboratory

P.O. Box 808, L-235

Livermore, California 94551 USA

[2]Georgetown University, Lombardi Comprehensive Cancer Center, 3970 Reservoir Road, NW, Washington, DC 20057 USA

Keywords: bio-signature, class prediction, canonical variate analysis, discriminant analysis, overfit, supersaturated designs

Running Head: Canonical analysis of groups in expression data

Author Footnote: J.R. Kercher and K.J. Wu are Physicist and Chemist, respectively, Lawrence Livermore National Laboratory, P.O. Box 808, L-235, Livermore, CA 94551 (emails: kercher1@llnl.gov and wu17@llnl.gov, respectively); J.N. Quong and A.A. Quong are Asst. Prof. And Assoc. Prof., respectively, Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, D.C. 20057 (emails: jnq@georgetown.edu and aaq@georgetown.edu, respectively).

ABSTRACT

Canonical analysis (CA) discriminates observations classified into groups or clusters. However, gene- and protein-expression datasets often have such large numbers of variables that the canonical equation is not directly soluble, and conventional stepwise procedures are prone to overfit. We propose a new method of CA for data classified into two or more groups for which the number of variables (plus the number of groups) exceeds the number of observations. First, we find the least-squares solution to the canonical equation by solving it exactly in the subspace given by the range of the within-group variance matrix, which we refer to as the subspace canonical analysis (SCA). We rank the variables based on the results of the SCA using one of seven different criteria, defined by sensitivities and correlations. We then use the highest ranked variables as input to a backward elimination CA (BECA). In an example using TOF-SIMS protein data, the misclassification rates for the seven criteria ranged from nine percent to 32 percent, which compare favorably to the 54 percent for the conventional BECA with a univariate filter. In this example, we also show that a test in the stepwise procedure could distinguish faulty data from real data. In a second example of micro-array measurements of gene expression for two tumor types, primary and metastases, we find that the new method outperforms the conventional method in class prediction for four of the six ranking criteria. Analyzing this data as seven groups (the metastases group and six groups of primary tumors classified by origin), the metastases and up to three of the six primary types, depending on the criterion, can be simultaneously discriminated.

INTRODUCTION

Kercher et al. (2004) developed a new two-stage method of canonical analysis (CA, Seal 1964) for the discrimination of two groups when $p>(N–h)$, which is common in gene- and protein-expression data sets, where p, h, and N are the number of variables (features), groups, and observations, respectively. We shall refer to their method as the *two-group* procedure. Canonical analysis is known under various names including canonical variate analysis (Krzanowski 2000) and discriminant coordinates (Seber 1984). When CA is applied to two groups it is also known as Fisher's linear discriminant analysis (FLDA). The two-group procedure was designed to reduce "overfit". In this paper we generalize the two-group procedure to an arbitrary number of groups and consider the important problems of class prediction, overfit, and signal detection in class discrimination.

The overfit problem. With the advent of modern analytical techniques, the number of measured variables in gene- or protein-expression experiments can range from tens to hundreds to thousands depending on the technology employed and the design of the experiment. Often the number of measured variables exceeds the number of observations that can be made with limited resources. Simon (2003), Simon et al. (2003), Dudoit et al. (2002), Radmacher et al. (2002), Nguyen and Rocke (2002), and Ambroise and MacLachlan (2002) point out that when p is greater than $(N–h)$, it typically leads to overfit for discrimination techniques. Overfit is a type of "capitalization on chance" (e.g., Harris 2001), in which a multivariate procedure exploits random fluctuations in a set of hundreds or thousands of variables to concoct a solution with apparently good discrimination, but which is unreliable outside the original data set.

The multiple-group problem. While gene- and protein-expression experiments frequently involve just two groups (e.g., malignant group/ benign group, treatment/control, etc), they can often involve more than two groups. For example a two-by-two factorial design has four groups; in drug studies there can be many treatment-dose levels; many types of tissue could be sampled in pathology studies; patients could be categorized by types or severity of infections in immunological studies, etc.

Class prediction problem and CA. An important problem in gene- and protein-expression studies is to predict group membership of an unknown observation given a data set of observations with known group membership. Typically one constructs a discrimination function based on the data of known memberships (training set) and uses this function to assign membership to unknown observations (test set). Dudoit et al. (2002) review several methods of discrimination for application to gene-expression studies. CA is a multivariate, supervised method for grouped data that maps observations with a large number of variables to a new space, usually of far fewer dimensions, in which the groups are optimally separated. The CA specifies which variables are important in discriminating between the groups and the significance of the separation of the groups. One can then classify unknown samples using the CA mappings.

The two-group procedure and its generalization to multiple groups. In the standard CA (Seal 1964, Krzanowski 2000), the number of variables p is less than or equal to $(N–h)$, $\mathbf{W}$ is nonsingular and $\mathbf{W}^{-1}$ exists where $\mathbf{W}$ is the within group sum-of-squares-and-cross-products (SSCP) matrix in the canonical equation. In this paper, we shall assume $p>(N–h)$, $\mathbf{W}$ is singular, and $\mathbf{W}^{-1}$ does not exist. The standard methods of CA depend on the nonsingularity of $\mathbf{W}$. However, the two-group procedure of Kercher et

al. (2004) analyzes the canonical equation for singular **W**. In the first stage of the two-group method, they convert the canonical (eigenvalue) equation to a linear system, which is possible in the two-group problem because of its special nature. They then find the minimal least squares solution using the Moore-Penrose generalized inverse (MPGI). The two-group procedure selects the ($N–h$) most important variables in the MPGI solution, and uses this shorter list as input to the second stage, which is a stepwise FLDA, either a forward selection (FS) or a BE algorithm.

Here we generalize the two-group procedure to apply to any number of groups, i.e., two or more. In the first stage, we find the least-squares solution to the canonical equation. This reduces to solving an eigenvalue problem exactly in the subspace defined by the range of **W**, which we refer to as the subspace canonical analysis (SCA). We use the SCA solution to rank the variables in importance. We truncate the list for input to the second stage, which is the stepwise CA as in the two-group procedure. In the Supplement Section I, we show that, for the case of two groups, the SCA solution is identical to the MPGI solution.

Contrast between the conventional method and the new method. In a conventional two-stage method for $p>(N–h)$, the first stage sorts the variable list with a univariate statistic such as the BW ratio of Dudoit et al. (2002). The variable list is truncated and then used as input to the discriminant procedure. In contrast, the method proposed here sorts the variables based on the results of the multivariate SCA, which take into account the inherent multivariate correlations among the variables. We propose seven different criteria to rank each variable's contribution to the SCA result.

Goals. In this paper we describe the SCA procedure and the new two-stage method. We apply it to two example data sets: one of protein data (four groups) and the other gene-expression data (two groups and seven groups). In both cases we shall apply the new method to class prediction and compare results to a conventional two-stage method. We also describe four new ranking criteria in addition to the three that were used in the two-group procedure.

EXPERIMENTAL SYSTEMS AND METHODS

Mass spectrometry measurements of fragments of pure proteins. The ability to identify the constituents of intracellular components would have far reaching implications on the study of cellular structure and function. As a step towards this goal, Quong et al. (2004) have investigated the use of Time-of-Flight, Secondary-Ion-Mass-Spectrometry (TOF-SIMS) to find signatures for specific proteins. Their data consists of four groups of 30, 30, 30, and 29 TOF-SIMS measurements on pure samples of cytochrome, insulin, lysozyme, and myoglobin, respectively, using a PHI THRIFT III instrument. We use mass-to-charge (M/Z) ratios 50 to 400 amu/(positive electric charge). The M/Z values from 0 to 49 and from 401 to 1000 were discarded.

Microarray gene expression data of primary and metastatic tumors. Ramaswamy et al. (2003) have analyzed data of primary tumors and metastases to study differences in their genetic expression. First, they found the 128 genes with the highest signal-to-noise ratio (S_x) between primary and metastatic tumors in Dataset A, which is Affymetrix Hu6800/Hu35KsubA-microarray data for 64 primary and 12 metastatic tumors from six different sites of origin (breast, lung, colorectal, prostate, uterus, and ovary). They investigated discrimination of primary tumors in other data sets using these 128 genes.

We will analyze the 128-gene data set for two groups, i.e., the 64 primary and 12 metastatic tumors. Using the transformation to canonical space found by the SCA-BECA, we will make class predictions for the primary lung tumors from their Dataset B and present the error rate. We will also analyze the 128-gene data set of the 64 primary and 12 metastatic tumors as a seven-group problem by dividing the primary tumors into six groups based on their site of origin. Finally we will analyze the data set of lung primary tumors divided into the two groups: recurring and non-recurring.

ALGORITHM

The Canonical Aquation

In CA, one assumes that the random variables are independent and identically distributed with a multivariate-normal distribution. See Suppl. Sec. I for a brief review of definitions and notation to be used. The transformation to the canonical space y for the *p*-vector of the *i*th observation $\mathbf{x}_s^i$ is given by

$$\mathbf{y}^i = \mathbf{E}^{\mathbf{T}}\mathbf{x}_s^i. \qquad (1)$$

where superscript **T** indicates the transpose. Let $\mathbf{e}^i$ be the *i*th column vector of **E**. Then the canonical equation is given by

$$(\mathbf{B} - \lambda_i \mathbf{W})\mathbf{e}^i = 0, \qquad (2)$$

where **B** is the between group SSCP. The transformations are subject to the normalization equations

$$\mathbf{e}^{i\mathbf{T}}\mathbf{W}\mathbf{e}^i/(N-h) = 1 \qquad (3)$$

Note that only (h–1) of the λ_i are nonzero. Transforming either $\mathbf{x}_s^i$ or $\left(\mathbf{x}_s^i - \bar{\mathbf{x}}_s\right)$ to the new space by eq. 1 produces either absolute or relative (grand-mean centered at 0) scales,

respectively, where $\bar{\mathbf{x}}_s$ is the *p*-vector sample grand mean. The latter is more convenient for exposition purposes; the former is more convenient to extrapolate to other data sets.

In this paper we assume that $p>(N-h)$. In which case, the $p\times p$ matrix $\mathbf{W}$ is singular, i.e., the rank r of $\mathbf{W}$ is bounded by $r \le (N-h)$, and almost always, $r=(N-h)$.

Subspace Analysis of the Canonical Equation for p>(n–h)

We first find the spectral decomposition of the symmetric matrix $\mathbf{W}$, $\mathbf{V}^{\mathrm{T}}\mathbf{W}\mathbf{V}=\mathbf{\Delta}$. The matrix $\mathbf{V}$ is nonsingular and orthogonal; and $\mathbf{\Delta}$ is the diagonal eigenvalue matrix of $\mathbf{W}$ where $\Delta_{ii}\ge\Delta_{jj}$ if $i<j$. Define the $r\times r$ matrix $\mathbf{\Delta}_{\mathbf{W}}$ such that $(\mathbf{\Delta}_{\mathbf{W}})_{ij}=(\mathbf{\Delta})_{ij}$ for $i,j\le r$. We define $\mathbf{R}_{\mathbf{W}}=\mathbf{\Delta}_{\mathbf{W}}^{-1}$. The $p\times p$ matrix $\mathbf{R}$ has elements $(\mathbf{R})_{ij}=(\mathbf{R}_{\mathbf{W}})_{ij}$ for $i,j\le r$ and $(\mathbf{R})_{ij}=0$ otherwise; $\mathbf{R}$ is the MPGI of $\mathbf{\Delta}$.

Either the range of $\mathbf{B}$ lies entirely in the null space of $\mathbf{W}$, $\Re(\mathbf{B})\cap\Re(\mathbf{W})=\varnothing$, or the range of $\mathbf{B}$ lies, at least partially, in the range of $\mathbf{W}$, $\Re(\mathbf{B})\cap\Re(\mathbf{W})\ne\varnothing$. We shall assume the latter in this paper. The point $\mathbf{B}\mathbf{e}^i$ lies in the range of $\mathbf{B}$ and the point $\lambda_i\mathbf{W}\mathbf{e}^i$ lies in the range of $\mathbf{W}$. Under the assumption $\Re(\mathbf{B})\cap\Re(\mathbf{W})\ne\varnothing$, we can project the point $\mathbf{B}\mathbf{e}^i$ orthogonally onto the range of $\mathbf{W}$ and set the projected point equal to $\lambda_i\mathbf{W}\mathbf{e}^i$. Note that $\mathbf{e}^i$ is then the least squares solution to eq. 2 because the sum of the squared residuals is the squared Euclidian distance between $\mathbf{B}\mathbf{e}^i$ and $\lambda_i\mathbf{W}\mathbf{e}^i$, which is minimized for this $\mathbf{e}^i$. Thus, eq. 2 becomes

$$\mathbf{W}\mathbf{W}^{+}\mathbf{B}\mathbf{e}^i=\lambda_i\mathbf{W}\mathbf{e}^i \qquad (4)$$

where $\mathbf{WW}^{+}$ is the orthogonal projection operator onto the range of $\mathbf{W}$ (e.g., Schott 1977 p. 173) and $\mathbf{W}^{+}$ is the MPGI of $\mathbf{W}$. Because $\mathbf{W}$ is symmetric, $\mathbf{W}^{+} = \mathbf{VRV}^{\mathbf{T}}$ (e.g., Schott 1997 p. 177). Eq. 4 equates two vectors in the r-dimensional space of the range of $\mathbf{W}$. The bottom (p–r) rows of eq. 4 are identically 0 on both sides of the equation. Thus, eq. 3 and eq. 4 together contain (p+1) unknowns and (r+1) equations for a vector $\mathbf{e}^{i}$ and scalar λ_i. The system is underdetermined with (p–r) indeterminate unknowns. The normalization equation is of the form

$$\mathbf{e}^{i\mathrm{T}}\mathbf{W}\mathbf{e}^{i} = \mathbf{e}^{i\mathrm{T}}\mathbf{V}\mathbf{V}^{\mathrm{T}}\mathbf{W}\mathbf{V}\mathbf{V}^{\mathrm{T}}\mathbf{e}^{i} = \mathbf{f}^{i\mathrm{T}}\Delta\mathbf{f}^{i} = \mathbf{f}_{\mathbf{r}}^{i\mathrm{T}}\Delta_{\mathbf{W}}\mathbf{f}_{\mathbf{r}}^{i} = N - h \qquad (5)$$

where $\mathbf{f}_{\mathbf{r}}^{i}$ is the first r components of the vector $\mathbf{f}^{i}$, which is defined as

$$\mathbf{f}^{i} = \mathbf{V}^{\mathrm{T}}\mathbf{e}^{i}, \qquad (6)$$

Also, define the vector $\mathbf{f}_{\mathbf{I}}^{i}$ as the last (p–r) components of $\mathbf{f}^{i}$, i.e., $\mathbf{f}^{i\mathrm{T}} = \left(\mathbf{f}_{\mathbf{r}}^{i\mathrm{T}}, \mathbf{f}_{\mathbf{I}}^{i\mathrm{T}}\right)$. Eq. 5 indicates that the normalization equation only fixes the length of $\mathbf{f}_{\mathbf{r}}$ with r components and not $\mathbf{f}_{\mathbf{I}}$ with (p–r) components. The projection of $\mathbf{e}^{i}$ onto the range and null space of $\mathbf{W}$ is given by $\mathbf{e}_{\mathbf{s}}^{i} = \mathbf{WW}^{+}\mathbf{e}^{i} = \mathbf{V}_{\mathbf{r}}\mathbf{f}_{\mathbf{r}}^{i}$ and by $\mathbf{e}_{\mathbf{I}}^{i} = \left(\mathbf{I} - \mathbf{WW}^{+}\right)\mathbf{e}^{i} = \mathbf{V}_{\mathbf{I}}\mathbf{f}_{\mathbf{I}}^{i}$, respectively, where the $p \times r$ matrix $\mathbf{V}_{\mathbf{r}}$ and the $p \times (p-r)$ matrix $\mathbf{V}_{\mathbf{I}}$ are the first r and the last (p–r) columns of $\mathbf{V}$, respectively. Furthermore, $\mathbf{f}_{\mathbf{I}}$ appears on the left side only of the first r rows of eq. 4, but $\mathbf{f}_{\mathbf{r}}$ appears on both sides. The component of $\mathbf{e}^{i}$ in the null space of $\mathbf{W}$, $\mathbf{e}_{\mathbf{I}}^{i}$, is indeterminate. If we transform eqs. 4 through 6 using $\mathbf{y}^{i\mathrm{T}} = \left(\left[\boldsymbol{\Delta}_{\mathbf{W}}^{1/2}\mathbf{f}_{\mathbf{r}}^{i}\right]^{\mathbf{T}} \quad \mathbf{f}_{\mathbf{I}}^{\mathbf{T}}\right)$, then

$\mathbf{y}_{\mathbf{W}}^{i\mathbf{T}} = \left(\left[\boldsymbol{\Delta}_{\mathbf{W}}^{1/2} \mathbf{f}_{\mathbf{W}}^{i} \right]^{\mathbf{T}} \quad \mathbf{0}^{\mathbf{T}} \right)$ is the minimal solution to the resulting canonical equation where $\mathbf{f}_{\mathbf{W}}^{i}$ is the exact solution to

$$\mathbf{R}_{\mathbf{W}} \mathbf{V}_{\mathbf{r}}^{\mathbf{T}} \mathbf{B} \mathbf{V}_{\mathbf{r}} \mathbf{f}_{\mathbf{W}}^{i} = \lambda_i \mathbf{f}_{\mathbf{W}}^{i} . \qquad (7)$$

So we restrict our solution to the projection of $\mathbf{e}^i$ onto the range of **W**, i.e., $\mathbf{e}_{\mathbf{s}}^{i} = \mathbf{W}\mathbf{W}^{+}\mathbf{e}^{i} = \mathbf{V}\boldsymbol{\Delta}\mathbf{R}\mathbf{f}^{i}$ or $\mathbf{e}_{\mathbf{s}}^{i\mathbf{T}} = \mathbf{f}^{i\mathbf{T}}\mathbf{V}^{\mathbf{T}} = \left(\mathbf{f}_{W}^{i\mathbf{T}}, \quad \mathbf{0} \right) \mathbf{V}^{\mathbf{T}}$. We see that $\mathbf{f}_{\mathbf{W}}^{i}$ is normalized by eq.5. The eigenvalue equation, eq. 7, has (h–1) nonzero eigenvalues λ_i and (r–h+1) zero eigenvalues. After finding $\mathbf{f}^i$, $\mathbf{e}^i$ is reconstituted using eq. 6. The (h–1) solutions of interest, $\mathbf{e}^i$, are all least squares solutions to eq. 2 and all lie in the range of **W**. In Suppl. Sec. II, we prove that if the range of **B** is a subset of the range of **W**, then the solutions of eq. 7 are exact solutions to eq. 2, the canonical equation. We also show that the residual vectors of eq. 2 all lie in the null space of **W**, and simplify eq. 7 for computation.

Statistical inference in the subspace method. The whole machinery of statistical inference may be applied to the subspace canonical analysis. See Suppl. Sec. III.

Statistical inference in the stepwise methods. See Suppl. Sec. III for implementations of the inference tests in the stepwise method, including definitions of *P*-values associated with the tests, such as $P_{\chi^2}(i)$ for the significance of the eigenvectors and $P_{FW}(i)$ for Wilks Λ for i variables remaining. We use the test of Rao (1970), Hawkins (1976), and McHenry (1978), which gives *P*-value $P_{Ft}(i)$ in rejecting the hypothesis that the ith eliminated variable does not significantly improve Wilks Λ. At each step in the BECA, we eliminate the least significant variable. Rao originally

developed this test for two groups; Hawkins and McHenry extended it to multiple groups. Hawkins (1976) recommended Bonferroni adjustments be applied to $P_{Ft}(i)$.

Implementation of the SCA-BECA Algorithm and Ranking Results of the Subspace Canonical Analysis for Stepwise Algorithms

The SCA finds the least squares solution to the canonical equation. These solutions almost always have all *p* coefficients of e^i be nonzero. However, the removal of most of the variables either does not materially change the solution or improves discrimination. Because the SCA solutions are the best solutions in the sense of least squares, our goal is to remove those variables contributing least to the SCA solution. Hence, we rank each variable by its contribution to the SCA solution, and then truncate the resulting list at the maximum number of variables for which the standard CA applies, namely (*N*–*h*). Then this shortened list of variables is input for a stepwise CA, either forward selection (FSCA) or backward elimination (BECA), which produces a final variable list and associated CA.

We designate the new procedure as SCA-X-BECA where X refers to one of seven different ranking criteria, which decides the importance of each variable to the SCA. Criteria 1, 5, and 6, listed below, were used in the two-group procedure. We introduce the other four criteria here. Implementation methods of the SCA-BECA algorithm and details of the criteria are in Suppl. Sec. IV.

Criterion 1: Sensitivity of the Wilks ratio as the between-group SSCP responds to changes in coefficients. This was designated as Criterion A in the two-group procedure.

Criterion 2: Eigenvector coefficient size. Krzanowski (2000) has suggested that the size of the coefficients in the eigenvector **e**, modified with the proper weighting to

correct for the normalization equation, can be used to determine the importance of each variable. Quong et al (2004) use the squares of the coefficients to weight the relative contribution of each original variable to the canonical variable. We combine the two notions to construct a ranking score for each variable in our second criterion.

Criterion 3: Probability of coefficients (*Significance of eliminated variable*) *and probability of Wilks ratio*. The probability of significance of variable *k* is based on the coefficients of $\mathbf{V_r}$, $P_{FW}(i–1)$ for the (i–1) variable set of the BE procedure, and $P_{Ft}(i)$ for the significance of the variable i . Here we assume that the probability for the ith variable is conditioned on the probability of significance of the (i–1)th set of variables.

Criterion 4: Cumulative probability of coefficients. This is similar to Criterion 3, except that we replace $P_{FW}(i–1)$ by the cumulative probability for (i–2) times $P_{Ft}(i–1)$.

Criterion 5: Absolute sensitivity. We sort on $|\Gamma_i|$ in this criterion, which is related to Criterion 1. This was referred to as Criterion B in the two-group procedure.

Criterion 6: Maximum correlation. For this criterion and the next, we assume that the correlation of the original variable with the significant canonical axes produced by the subspace CA determines its importance. In Criterion 6, we set a significance limit of α. If the canonical axis has $P_{\chi^2}(i) < \alpha$, then it is included. The correlation of variable *k* is found for each of these significant axes and the maximum correlation is used.

Criterion 7: Total correlation. This ranking criterion differs from Criterion 6 only in that we use the correlation summed across all the canonical axes weighted by the estimated probability for that axis. Note that for the case of two groups, Criterion 6 and

Criterion 7 are identical. These two criteria were referred to as Criterion C in the two-group procedure.

*Conventional two-stage canonical analysis***.** The first stage of the conventional two-stage CA is a univariate ranking and truncation algorithm. After sorting the variables on the univariate BW ratio of Dudoit et al. (2002) in descending order, we select the first ($N–h$) variables for input to the second stage, which is a BECA. See Suppl. Sec. IV.

RESULTS AND ANALYSIS

Mass Spectrometry Measurements of Fragments of Four Pure Proteins.

Parameters for canonical analysis. In this exercise, the number of measurements is N=119, the number of groups is h=4, the number of variables is p=351, and the size of the groups are $n_1=n_2=n_3=30$ and $n_4=29$.

Safeguards against analyzing noise. The CA algorithms are very effective in discriminating groups, especially when many variables are available from which the algorithms may select. Thus, one might question whether or not groups are being artificially discriminated in data, which is essentially noise. We investigate this problem using the TOF-SIMS protein data. In Fig. 1a, we show the positions in canonical space resulting from the SCA of the four types of proteins. Canonical space axes, defined by eq. 1, are ordered by the size of the corresponding eigenvalues. Canonical axis 1 separates myoglobin at one extreme and cytochrome at the other. Insulin and lysozyme are not distinguished on axis 1. On canonical axis 2, insulin, lysozyme, myoglobin, and cytochrome lie along a gradient at relatively evenly spaced intervals.

Using this data, we make four artificial random groups. Group 1, group 2, group 3, and group 4 consists of observations of cytochrome, insulin, lysozyme, and myoglobin in the ratios of eight, eight, seven, seven; seven, seven, eight , eight; eight, seven, eight , seven; and seven, eight, seven, and seven, respectively. In Fig. 1b, we show the positions from the SCA of the random groups drawn on the same scale as Fig. 1a. The random groups are very tightly bunched together compared to the actual groups. This is a very encouraging result. However if we plot the random groups at a finer scale of resolution, as shown in Fig 1c, then a casual observer, not knowing that the groups were artificially generated, might think Fig. 1c indicates that random group 2 and random group 4 are separated on canonical axis 1 due to some real phenomenon. We suggest there is a safeguard to prevent one from drawing such a conclusion. In Fig. 2 we show the *P*-value $P_{Ft}(i)$ for rejecting the null hypothesis that the eliminated variable has not significantly improved Wilks Λ, i.e., the test due to Rao (1970), Hawkins (1976), and McHenry (1978). In Figs. 2a and 2b we show this *P*-value for SCA-1-BECA for the original data groups and for the randomized groups, respectively; in Figs. 2c and 2d, for the conventional method, BW-BECA. In the original data case, we find very high significance (very low *P*-values) for about the first 20 or so variables, unlike the random group case in which we find large values (and large fluctuations) of the *P*-value $P_{Ft}(i)$ even for very small numbers of remaining variables. Similar results were obtained for all criteria as shown in Suppl. Sec. V. In Fig. 2a, note that, as variables are eliminated, eventually there is an *i* such that $P_{Ft}(i) < \alpha$ where α is some pre-chosen critical value and for which $P_{Ft}(j) < \alpha$ for all $j<i$. We use α=0.05. We refer to *i* as the *cutoff* variable. We

also refer to *low-end* variables as those remaining variables *k* for which $P_{Ft}(k) < \alpha$ and *k* just above *i*, i.e., *k* within about ten units or so of *i*.

The canonical spaces. We show the position of the observations in the first two dimensions in canonical space produced by the ranking from Criteria 1 (Sensitivity of Wilks Λ) in Figs. 3a, 3b, 3c, and 3d for three, six, nine, and twelve fragments, respectively, used in the CA. Twelve variables is the Bonferroni-adjusted cutoff variable for SCA-1-BECA for this data. The average Bonferroni-adjusted cutoff for all criteria is fourteen variables. As fragments are added to the CA, the separation of the groups increases in canonical space. In Suppl. Sec. VI we show similar results when using Criteria 6 (Maximum correlation) for ranking. The plots for SCA-1-BECA are generally similar to those for SCA-6-BECA, but they differ in some details. For example, SCA-1-BECA separates the insulin cluster from the lysozyme better than SCA-6-BECA.

A superficial examination of the variables used to construct the spaces for the seven criteria suggests that differences between the spaces are pronounced (Suppl. Sec. VII). However if one examines the correlations of each measured variable with the canonical axes, one finds that the canonical spaces produced by the different criteria are remarkably similar (Suppl. Sec. VIII). Based on the rank correlations of variables with canonical axes, we conclude that Criteria 1 and 2 tend to produce spaces most similar to the other spaces; followed by Criteria 4 and 6; then by Criteria 3 and 5 and BW-BECA; and finally by Criterion 7 (Suppl. Sec. IX). This result is the same for both axes. Many variables (i.e., 307, 331, 223, 315, 399, 398, and 355 M/Z), which are not used to generate the axes, have higher correlations with the axes than the variables in the $\mathbf{e}^i$.

In another test, we examine the 14-fragment CA for each of the seven criteria and the conventional BW-BECA and compare the number of variables unique to that CA, i.e., not shared with any of the 14-fragment CA's of the other criteria (Suppl. Sec. X). We find that Criteria 1 and 2 produce lists of fragments that are most like the other lists, followed by Criteria 3 and 4, Criteria 5 and 6, Criterion 7, and BW-BECA.

Class prediction in TOF-SIMS protein data. Four months after the initial TOF-SIMS experiment, which produced the training-set data shown in Fig. 1a and Fig. 3, we conducted another TOF-SIMS experiment to produce an additional data set (test set). The test set had 20 observations (spectra) each for cytochrome, lysozyme, and insulin and 18 observations for myoglobin. We assigned group membership for the 78 test-set observations using the $\mathbf{e}^i$'s found with the training set. For multivariate, identically normally-distributed data, this assignment algorithm is equivalent to multivariate discrimination analysis using the Mahalanobis distance. We show the results of this class prediction in Table 1 along with the Wilks Λ statistics for class discrimination for Bonferroni-adjusted cutoff CA's. Criteria 1, 2 and 4 had a relatively low number of misclassifications. The conventional BECA had the most, with over a 50% error rate. The conventional method also produced the worst Wilks ratio at Bonferroni-adjusted cutoff. Criterion 4 had the best Wilks ratio and the second best *P*-value for the Wilks ratio test. The criterion that achieved the best classification error rate with the lowest number of variables used was Criterion 1: (Sensitivity of Wilks ratio). We show the actual mappings to canonical space for both the training set and test set for SCA-1-BECA, SCA-2-BECA, SCA-4-BECA, and BW-BECA (conventional) in Figs. 4a, 4b, 4c, and 4d, respectively. The other four CA's of Table 1 are shown in Suppl. Sec. XI. We

suggest that to construct a robust predictor, one should use a variety of runs over a wide range of experimental conditions that cover the status of the reagents and the use of the instruments. We also suggest that, despite the evident scatter in the test set, the method itself shows good promise for producing reliable predictions of class membership.

Microarray Gene Expression Data of Primary and Metastatic Tumors.

Class prediction by SCA/BECA for two groups. We test class prediction by the SCA-X-BECA method for micro-array data of primary tumors and metastases. We analyze the data set of Ramaswamy et al. (2003) comprising 128 genes from their Dataset A with highest signal-to-noise ratios for the two groups of 64 primary tumors and 12 metastases. In Table 2, we show the class predictions of the e1 mapping from the CA's of the 128-gene data set applied to the lung data set. The lung data set (Dataset B) consists of the gene expression on the Affymetrix U95A microarray for 62 primary lung tumors: 31 recurring and 31 non-recurring. Table 2 indicates that, for the Bonferroni-adjusted cutoff, Criteria 1, 2, 4, 5, and 6 provide excellent class prediction and are superior to the conventional BW-BECA. The class predictions of Criteria 1, 4, 5, and 6 either meet or exceed those for the conventional method (BW-BECA) for cutoff and low-end variables. Both SCA-3-BECA and BW-BECA are inferior to the other criteria. The performance of Criterion 2 is very sensitive in this instance to the number of variables. In Figs. 5 and 6 we show the boxplots for the training set and test set data for SCA-X-BECA for Criteria 2, 4, and 5 and for the BW-BECA for the Bonferroni-adjusted cutoff and low-end CA's, respectively. Figures for additional criteria are in Suppl. Sec. XII. On closer examination, we find that the gene AA010619 dominates the results for the lung primary tumor in Fig. 6a. The expression values for this gene were different by an

order of magnitude between the primary tumor data of the 128-gene data set and the lung primary tumor data. This change was coupled with a large coefficient in $\mathbf{e}^1$ for this gene to produce a large effect along the canonical axis. Criterion 2 was the only criterion for which this gene appears in the variable list used in the CA's of Table 2. In Suppl. Sec. XIII we give the transformations for the CA's listed in Table 2.

Canonical axis for two groups. We characterize the canonical axis by the variables that are highly correlated with it. In the Suppl. Sec. XIV, we give the 10 and 15 genes with the highest positive and negative correlation, respectively, with the canonical axis for each of the criteria. The same genes tend to show up on all or most of the lists and in similar positions. We found a similar result for the protein example

Effect of increasing variable count on predictability. Kercher et al. (2004) found for two groups that increasing the variable count within the same criterion usually improves the group separation-group size ratio. But as others point out (e.g., Dudoit et al. 2002), predictability is not stable under increases in the variable count. In all cases, which we have investigated to date, as we increase the number of variables beyond the low-end region, predictability degrades. We show examples in Suppl. Sec. XV.

Analysis of multiple groups: Seven groups. As an exploratory exercise, we analyze the 128-gene data set for separation between the different types of primary tumors and metastatic tumors. Recall that the genes were selected on the basis of the highest signal-to-noise ratios S_x separating metastases from primary tumors. Here our narrow goal is to see if the genes that best discriminate primary tumors from metastases also discriminate between primary types. In Fig. 7 we show the result of the subspace CA. Breast and lung tumors are not distinguishable in this CA. Prostatic tumors and

metastases are distinctly isolated from other groups. The remaining groups are in varying degrees of isolation and overlap. When we use low-end or cutoff CA's from the SCA-X-BECA, we often find that lung tumors and breast tumors are in proximity or co-located on the first canonical axis; the metastases and either the prostatic or uterine tumors are isolated. In Suppl. Sec. XVI we show $P_{Ft}(i)$ from the test of Rao (1970), Hawkins (1976), and McHenry (1978) against the number of remaining variables *i*, an example of the tests for group separation on canonical axes, and tests of significance of eigenvalues and eigenvectors in a BE case.

Lung data. Differences in two groups defined by recurrent/non-recurrent tumors. The 169-gene lung tumor dataset of Ramaswamy et al. (2003) is composed of two groups of 31 patients each; those with non-recurring and those with recurring tumors. In Suppl. Sec. XVII we show the results for a nine-gene and 18-gene CA from SCA-5-BECA. While these results are suggestive, we are reluctant to accept them as indicative of a signature for recurring/non-recurring tumors. The graphs of $P_{Ft}(i)$ as a function of *i* (Suppl. Sec. XVII) show a similarity to Fig. 2b and Fig. 2d, which are known to be for random groups. The graph of $P_{Ft}(i)$ for Criterion 5 is the one most suggestive of a real effect. If a true gene-expression difference exists between the recurring and non-recurring cases, it may be that the appropriate gene-expressions are not included in the data set. One possible approach to explore this is to filter Data set B on the signal-to-noise ratio for the recurring/non-recurring groups and investigate those genes. This exercise is beyond the scope of this paper. These results are shown as an example in which the methods in this paper may indicate that further analysis is required before one accepts a signature.

DISCUSSION

Use of the canonical analysis. When using a standard CA (nonsingular **W**), one usually wants to know how many significant dimensions are there in the new space; which groups are significantly separated in the new space; whether the groups are purely random; and what the biological or physical meanings of the axes are. To help answer these questions, it is pertinent to know what genes or proteins were used to generate the new space, and which genes or proteins are important. We have outlined a set of tests and procedures to answer these questions. Other important questions include whether the results are reproducible; whether similar data sets produce similar spaces; and whether the results are predictive. The answers to these questions may require additional data to test and compare results. In the applications shown above, we have suggested possible approaches to these questions and show how they can work in various circumstances.

In the nonstandard (singular **W**) CA, subset selection from the extensive list of available variables, which avoids overfit, is a central problem. We suggest that the method proposed here is a promising approach for variable selection in the discrimination and class prediction of gene- and protein-expression data. We can either regard the SCA as a filter for the input to the BECA or view the BECA as a refinement of the SCA. In either case, the SCA is the best solution in the sense of least-squares, and sorting on the results of the SCA takes into account the multivariate correlations among the variables, unlike the conventional BW-BECA, which just sorts on the univariate statistic BW.

Inferred properties of the canonical spaces; randomness and predictability. We can draw some general inferences from the analyses shown here, subject to further confirmation. First, we suggest that if the Rao-Hawkins-McHenry test for significance of

the eliminated variable is such that extremely low *P*-values are observed with few remaining variables, then this is an indication that the grouped data is analyzable with the variables measured in the data set. Second, the BE procedure following the subspace CA produces similar canonical spaces at the cutoff or low-end variables for all ranking criteria. Third, the use of Bonferroni-adjusted cutoff variables is protection for overfit. Using cutoff or low-end variables is frequently safe, but more risk is entailed than with Bonferroni-adjusted cutoff CA's. The SCA-X-BECA results for class prediction for Criteria 1, 4, 5, and 6 are superior to those for the conventional BW-BECA. For the Bonferroni-adjusted cutoff CA, we find that SCA-2-BECA is also superior to the conventional BW-BECA. Criterion 2 is suspect for CA's beyond the Bonferroni-adjusted cutoff. Criteria 3 and 7 are not satisfactory. We suggest these results are obtained because the SCA is the least squares solution to the canonical equation.

Remarks on the examples. The two data sets are very different in their properties. The groups in the data set of pure proteins produced by TOF-SIMS are well resolved with fewer variables than the groups in the gene data. We find it interesting that so few protein fragments can so decisively separate the four types. The new two-stage SCA-BECA method proposed here shows promise for constructing discrimination functions to classify protein samples.

The primary lung tumors are well predicted to be primary tumors. The variables in the 128-gene data set are insufficient to discriminate all seven groups of tumors simultaneously if we group the primary tumors by sites of origin. The metastases group is discriminated and usually one or two of the primary types, but not all six primary types simultaneously. The Rao-Hawkins-McHenry test in the BE procedure indicates that

further analysis of the two groups of lung tumors, designated as recurrent or non-recurrent, is required.

Further work. There are many problems attendant to the study of CA of gene- and protein-expression data beyond the scope of the effort reported here that should be examined. For example, an area of considerable importance is the application of CA or a similar procedure to repeated measurements. This experimental design is quite common when following the course of a disease. Additional work must be done to apply CA techniques to this design.

In the context of the examples discussed here much further work could be done. For the protein case, testing on actual cell components should be a very interesting set of exercises. In the gene-data set example, Ramaswamy et al. (2003) had additional data sets that could be investigated by the SCA-stepwise procedure for a more complete picture of metastases/primary tumor discrimination or primary-type discrimination. Another interesting exercise would be to use the 17 genes, which they propose as a signature, in a standard CA on the original primary/metastases data set. Also, the SCA-BECA method could be applied to examine the differences between normal tissue and tumors. To avoid differences between normal tissue types confounding differences due to cancer processes, pre-filtering the data for those genes associated with cancer processes might be useful.

ACKNOWLEDGMENTS

We would like to thank members of the Pathomics project at Lawrence Livermore National Laboratory for invaluable discussions, especially Drs. B. A. Sokhansanj and J. B. Pesavento. This work was performed under the auspices of the U.S. Department of

Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48. This work was supported by the Pathomics project under the LLNL Laboratory Directed Research and Development program.

SUPPLEMENT

J. R.Kercher, J.R., J. N. Quong, K. J. Wu, A. A. Quong. Supplement to variable selection in canonical analysis of gene- and protein-expression data: the general case for multiple groups. http://{To be determined}.

REFERENCES

Ambroise, C., G.J. McLachlan. 2002. Selection bias in gene expression on the basis of microarray gene-expression data. Proc. National Acad. Sciences 99:6562-6566.

Dudoit, S., J. Fridlyand, T.P. Speed. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Statistic. Assoc. 97:77-87.

Harris, R.J. 2001. A primer of multivariate statistics. 3rd ed. Lawrence Erlbaum Assoc.:Mahwah, NJ.

Hawkins, D.M. 1976. The subset problem in multivariate analysis of variance. J. Royal Statist. Soc. B 38:132-139.

Kercher, J.R., R.G. Langlois, B.A. Sokhansanj, C.F. Melius, J.N. Quong, F.P. Milanovich, B.W. Colston, Jr., K.W. Turteltaub, A.A. Quong. 2004. Variable selection in canonical analysis of gene- and protein-expression data: the special case of two groups. (Submitted).

Krzanowski, W.J. 2000. Principles of multivariate analysis. Clarendon Press, Oxford.

McHenry, C.E. 1978. Computation of a best subset in multivariate analysis. Appl. Statist. 27:291-296.

Nguyen, D.V., D.M. Rocke. 2002. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics 18:1216-1226.

Quong, J.N., K.J. Wu, J.R. Kercher, M. Knize, K. Kulp, A.A. Quong. 2004. A signature-based method to distinguish time-of-flight secondary ion mass spectra from biological samples. (Manuscript to be submitted).

Radmacher, M.D., L.M. McShane, R. Simon. 2002. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 9:505-511

Ramaswamy, S., K.N. Ross, E.S. Lander, T.R. Golub. 2003. A molecular signature of metastasis in primary solid tumors. Nature Geneteics 33:49-54.

Rao, C.R. 1970. Inference on discriminant function coefficients. P. 587-602. *In* Essays on Probability and Statistics (R.C. Bose, I.M. Chakravarti, P.C. Mahalanobis, C.R. Rao, K.J.C. Smith, eds) University of North Carolina Press:Chapel Hill, NC.

Schott, J.R. 1997. Matrix analysis for statistics. John Wiley & Sons : New York.

Seal, H. 1964. Multivariate statistical analysis for biologists. Wiley.

Seber, G.A.F. 1984. Multivariate observations. John Wiley & Sons.

Simon, R. 2003. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. British Journal of Cancer 89:1599-1604.

Simon, R., M.D. Radmacher, K. Dobbin, L.M. McShane. 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classfication. Journal of the National Cancer Institute 95:14-18.

Table 1. Statistics of the new canonical analysis method for seven sorting criteria and the conventional stepwise method. These CA's occurred at cutoff in the Bonferroni-adjusted critical value in the test by Rao (1970), Hawkins (1976), and McHenry (1978) for the TOFS-SIMS analysis of four proteins (cytochrome, insulin, lysozyme, and myoglobin). The training set consisted of 119 observations for the four proteins of 30, 30, 30, and 29 each, and the test set was 78 observations of 20, 20, 20, and 18.

Analysis Method	Wilks ratio Λ	Rank of P_{FW}	Number of variables p	Number of classification errors
SCA-1-BECA	0.000218	3	12	11
SCA-2-BECA	0.000225	6	15	7
SCA-3-BECA	0.000158	4	15	24
SCA-4-BECA	0.0000491	2	18	11
SCA-5-BECA	0.0000577	1	16	25
SCA-6-BECA	0.000615	7	12	20
SCA-7-BECA	0.000283	5	13	22
BW-BECA (conventional)	0.000850	8	12	42

Table 2. Statistics of procedures for 128-gene data set of primary tumors and metastases from Ramaswamy et al. (2003) for which canonical transformations were calculated. Classification errors when applied to 169-gene data set of lung primary tumors are shown in column 6. For the class prediction test, the training set was 74 primary tumors of from six sites of origin and 12 metastases; the test set was 62 primary lung tumors.

Procedure/ Criterion	Largest eigenvalue λ_1	Wilks ratio Λ	Rank of P_{FW}	Number of variables	Number of classification errors
Cutoff variable for Bonferroni-adjusted critical values for P_{Ft}					
SCA-1-BECA	1.19	0.456	14	3	0
SCA-2-BECA	1.52	0.397	7	2	0
SCA-3-BECA	0.533	0.652	18	2	2
SCA-4-BECA	1.15	0.464	12	2	1
SCA-5-BECA	1.73	0.366	5	3	0
SCA-6,7-BECA	1.52	0.397	7	2	0
BW-BECA (conventional	1.00	0.500	15	3	2
Cutoff variable					
SCA-1-BECA	1.58	0.387	10	5	0
SCA-2-BECA	2.02	0.331	2	4	62
SCA-3-BECA	0.635	0.612	17	3	3
SCA-4-BECA	1.79	0.358	8	4	0
SCA-5-BECA	1.73	0.366	5	3	0
SCA-6,7-BECA	1.81	0.356	4	3	0
BW-BECA (conventional	1.34	0.428	13	4	0
Low-end variable					
SCA-1-BECA	NA	NA	NA	NA	NA
SCA-2-BECA	3.63	0.216	1	10	62
SCA-3-BECA	1.53	0.394	16	9	6
SCA-4-BECA	3.19	0.239	9	13	0
SCA-5-BECA	3.27	0.234	3	11	0
SCA-6,7-BECA	2.49	0.286	6	8	2
BW-BECA (conventional	2.53	0.283	11	12	5

Figure Captions

Fig.1. (a) Scatterplot of TOF-SIMS data in the first two axes of canonical space as found by the SCA. (b) Scatterplot of first two canonical axes found by subspace canonical anslysis for randomized groups using the same TOF-SIMS data as in Fig 1a. Plots (a) and (b) are shown at the scale of canonical units. (c) Same data as in Fig. 1b, plotted at finer scale.

Fig. 2. Graphs of the *P*-values $P_{Ft}(i)$ for the null hypothesis that the eliminated protein fragment does not improve the group separation plotted against the number of protein fragments remaining in the set used for the standard canonical analysis. We use the calculation of Rao (1970), Hawkins (9176), and McHenry (1978). We show results for actual data in (a) and (c), randomized data in (b) and (d). (a) and (b) are for Criterion 1: Absolute sensitivity of Wilks ratio in canonical space. (c) and (d) are for the conventional CA pre-filtered using BW ratio: BW-BECA..

Fig. 3. Scatterplots of clusters of cytochrome, insulin, lysozyme, and myoglobin fragment data in canonical spaces generated by BE using Criterion 1 (Sensitivity of Wilks ratio) to sort fragments for inclusion in the BE procudure. Sensitivities are calculated based on transformations from SCA. Canonical analyses used three, six, nine, and twelve fragments in (a), (b), (c), and (d), respectively.

Fig. 4. Projection of training set (open symbols) and test set (filled symbols) for four proteins analyzed by TOF-SIMS for Bonferroni-adjusted cutoff CA's. See Table 1. (a) SCA-1-BECA 12 fragments, (b) SCA-2-BECA 15 fragments, (c) SCA-4-BECA 18 fragments, and (d) BW-BECA 12 fragments.

Fig. 5. Boxplots comparing original data set of primary tumors and metastases groups (training set) and 169-gene lung primary tumor data (test set) from Dataset B on the same axis in canonical space. Data from Ramaswamy et al. (2003). The lung primary data was projected onto the axis using eq. 1 where the transformation matrix was found using the primary tumors and metastases data from the 128-gene data set. The top and bottom of a box is the 75^{th} and 25^{th} percentile respectively; the horizontal line inside the box is the median. Any datum a distance away from either the top or the bottom greater than 1.5 times the distance between top and bottom is an outlier and shown by an open circle. The range of the remaining data is shown by horizontal bars above and below the box. These CA's are for Bonferroni-adjusted cutoffs. (a) SCA-2-BECA, two genes, (b) SCA-4-BECA, two genes, (c) SCA-5-BECA three genes, (d) BW-BECA, three genes

Fig. 6. As in Fig. 5, 128-gene primary tumor and metastases data was used as a training set. Transformations from training set CA's used to project 169-gene lung primary tumor data onto canonical axis. Data from Ramaswamy et al. (2003). (a) SCA-2-BECA 10 gene, (b) SCA-4-BECA 13 gene, (c) SCA-5-BECA 11 gene, (d) BW-BECA 12 gene.

Fig. 7. Subspace canonical analysis of 128-gene data set divided into seven groups: six primary tumors (breast, colorectal, lung, ovary, prostate, uterus) and one metastases group.

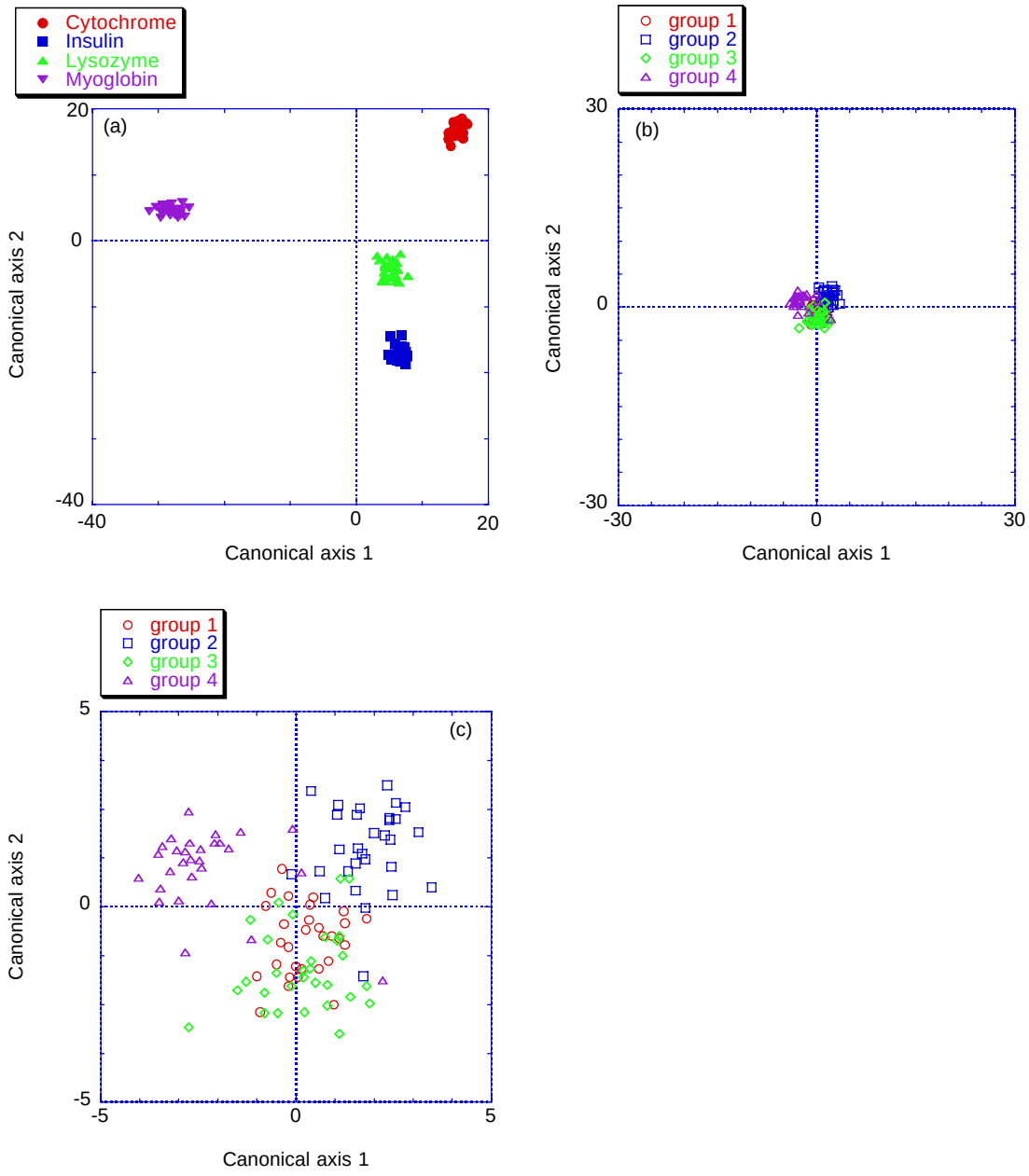

Cytochrome
Insulin
Lysozyme
Myoglobin
(a)
Canonical axis 2
Canonical axis 1
20
0
-40
-40
0
20
group 1
group 2
group 3
group 4
(b)
Canonical axis 2
Canonical axis 1
30
0
-30
-30
0
30
group 1
group 2
group 3
group 4
(c)
Canonical axis 2
Canonical axis 1
5
0
-5
-5
0
5

Fig. 1

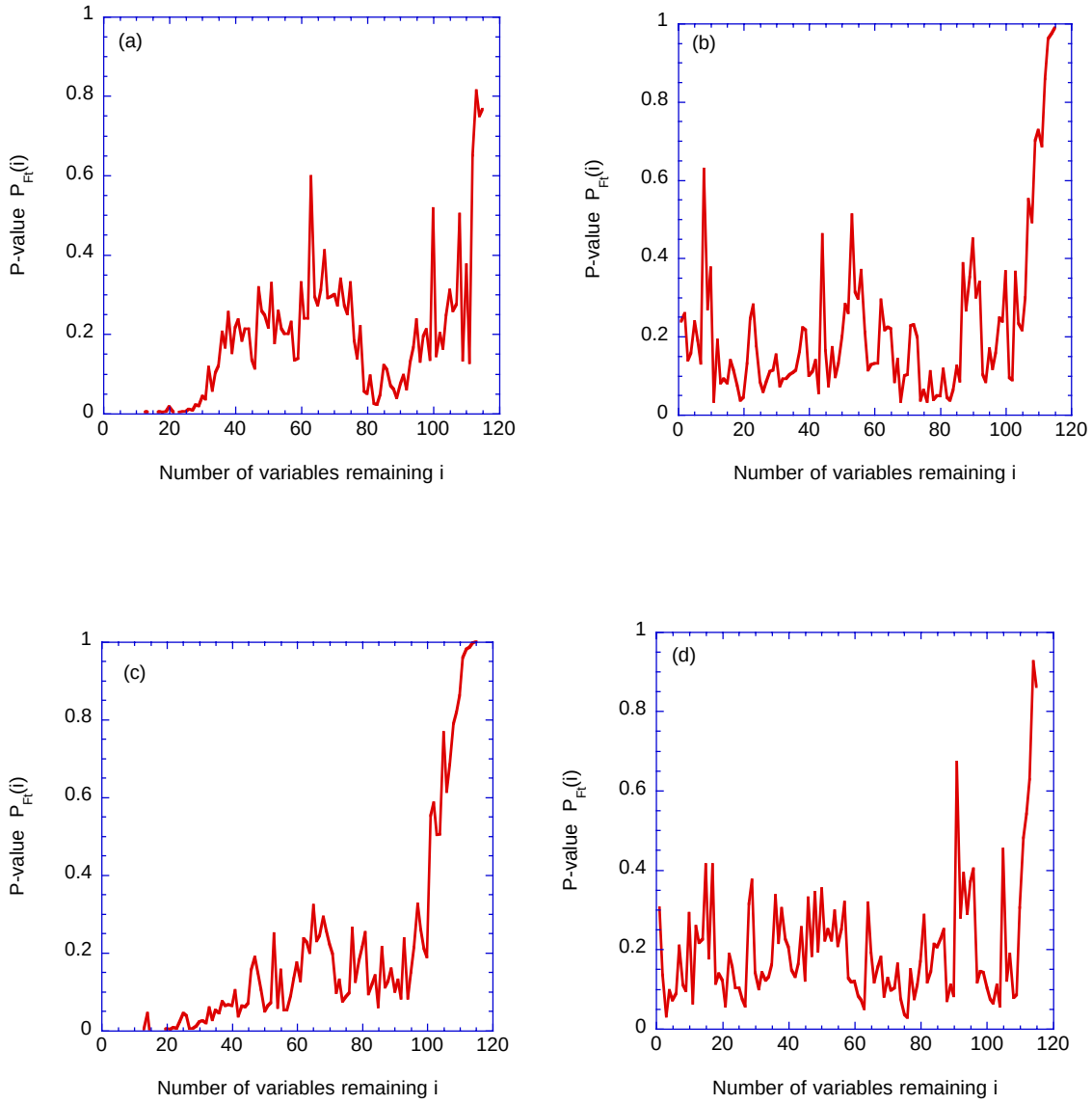

(a)
(b)
(c)
(d)
P-value $P_{Ft}(i)$
Number of variables remaining i
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100
120

Fig. 2

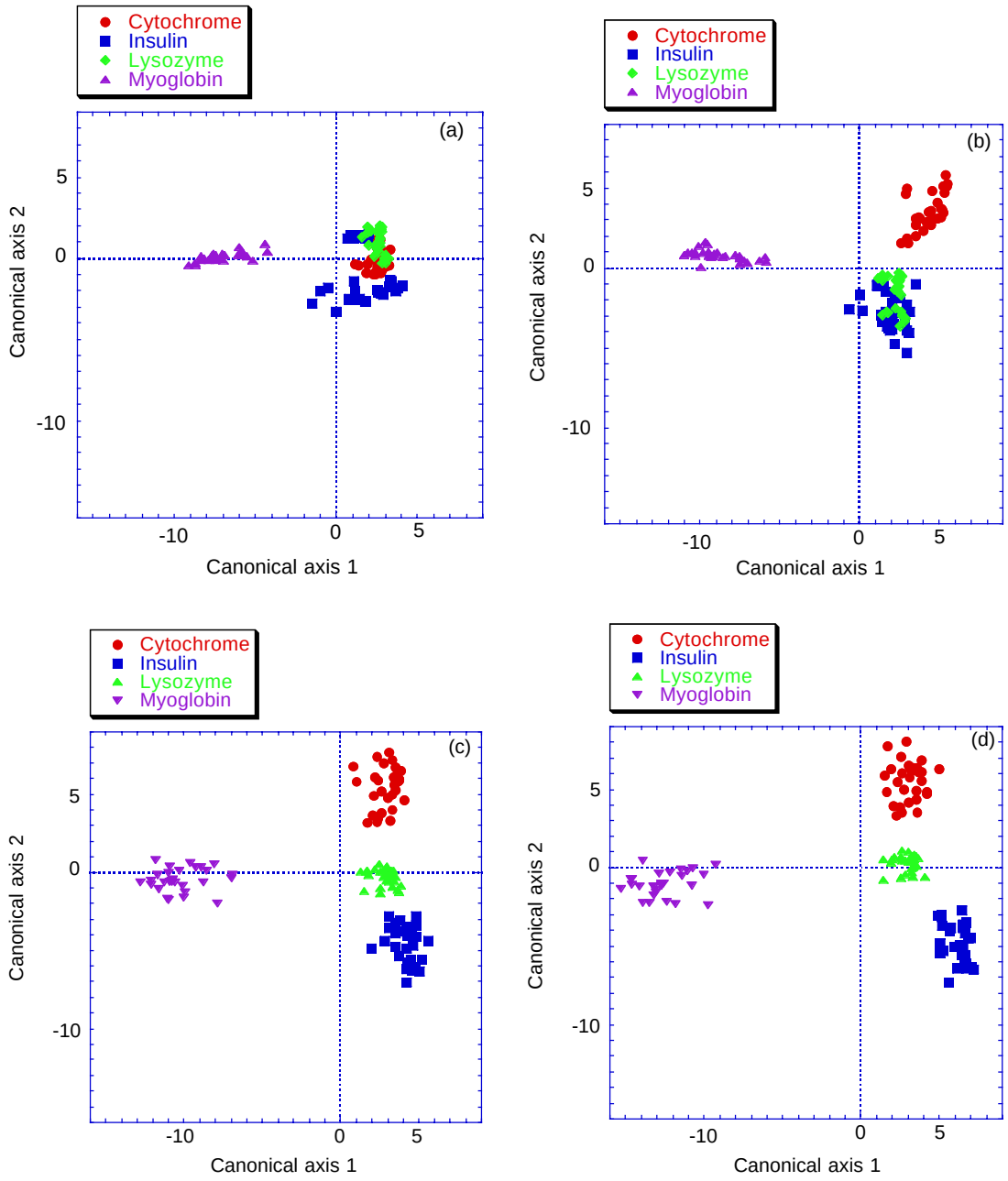

Cytochrome
Insulin
Lysozyme
Myoglobin
(a)
(b)
(c)
(d)
Canonical axis 1
Canonical axis 2
5
0
-10

Fig. 3

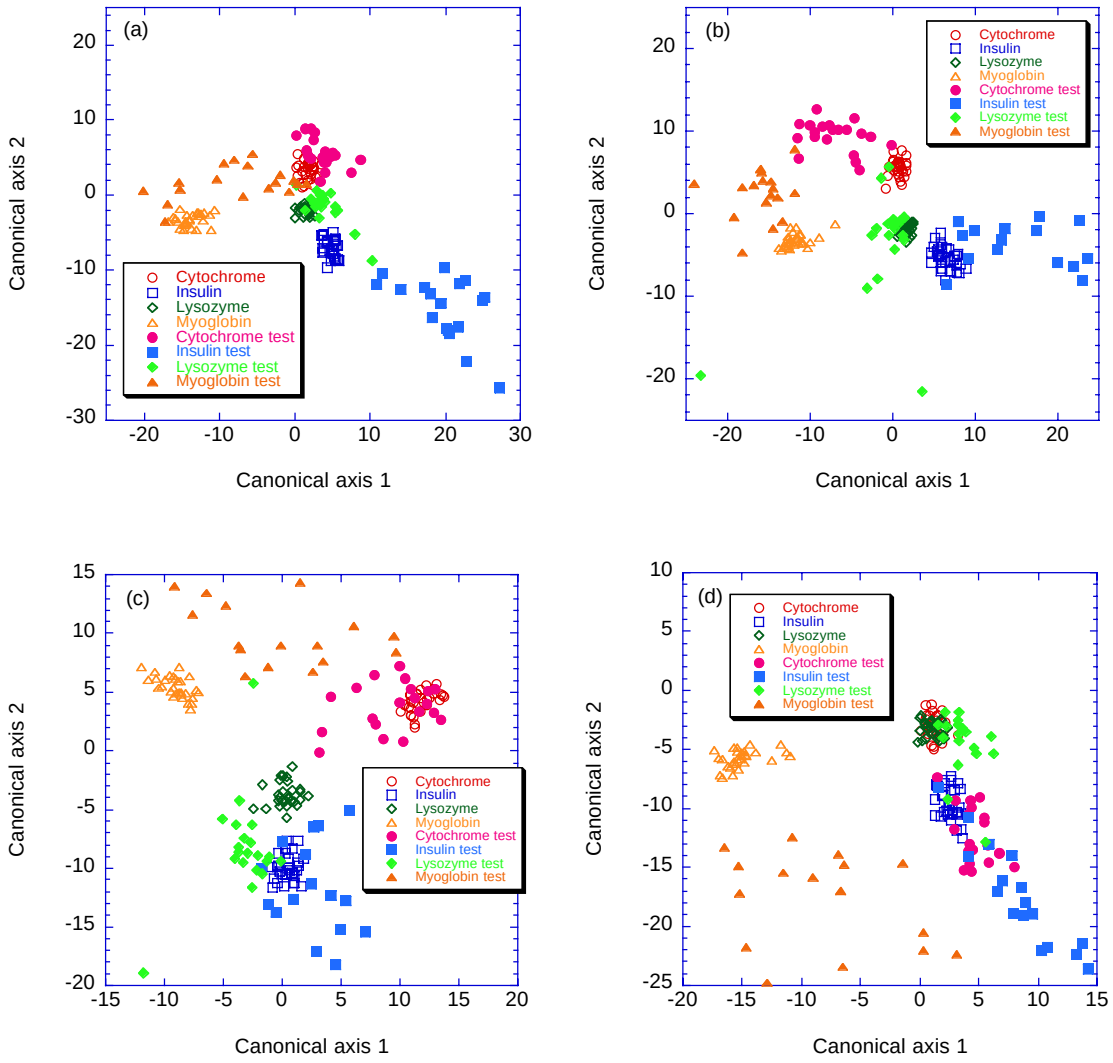

(a)
(b)
(c)
(d)
Cytochrome
Insulin
Lysozyme
Myoglobin
Cytochrome test
Insulin test
Lysozyme test
Myoglobin test
Canonical axis 1
Canonical axis 2

Fig 4

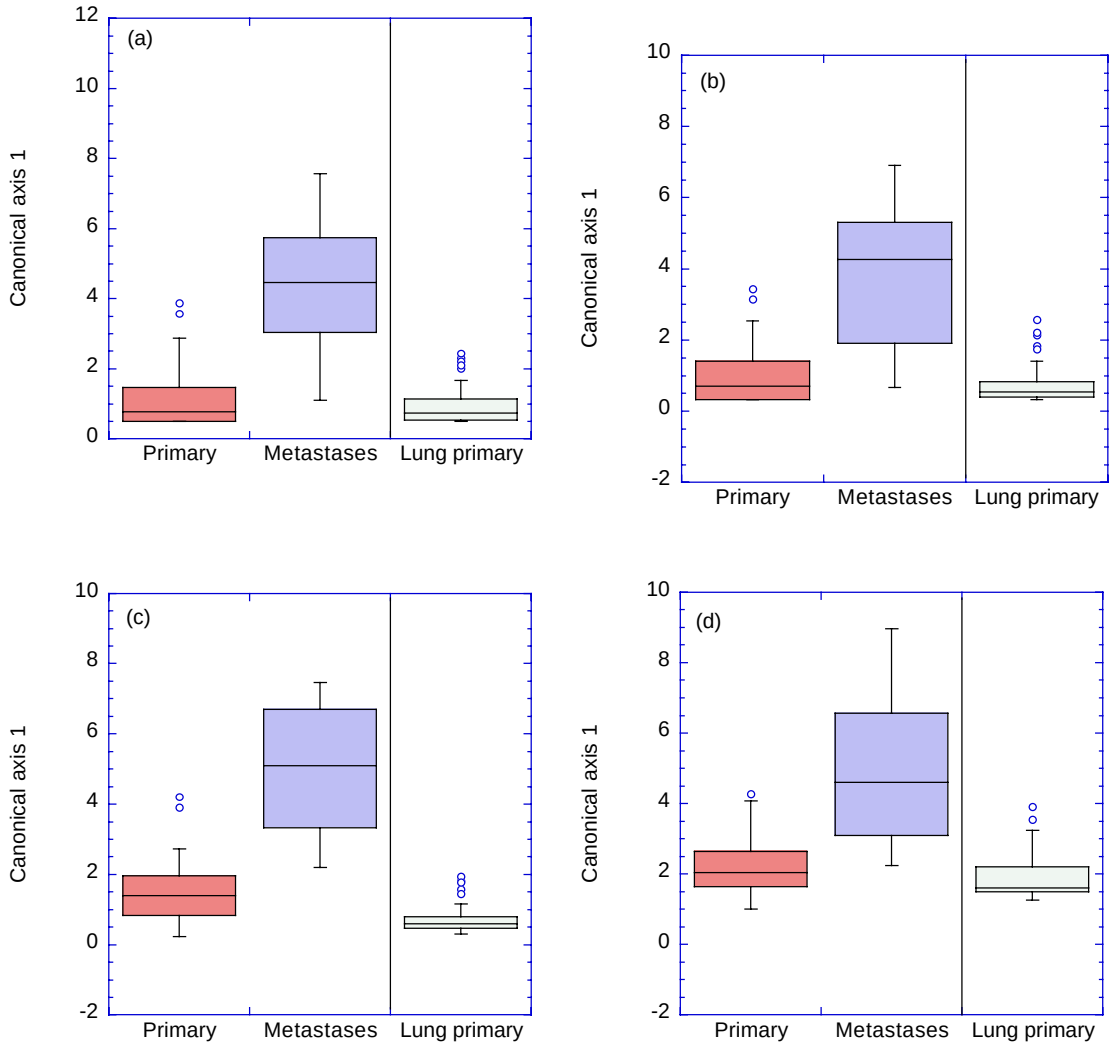


Fig. 5

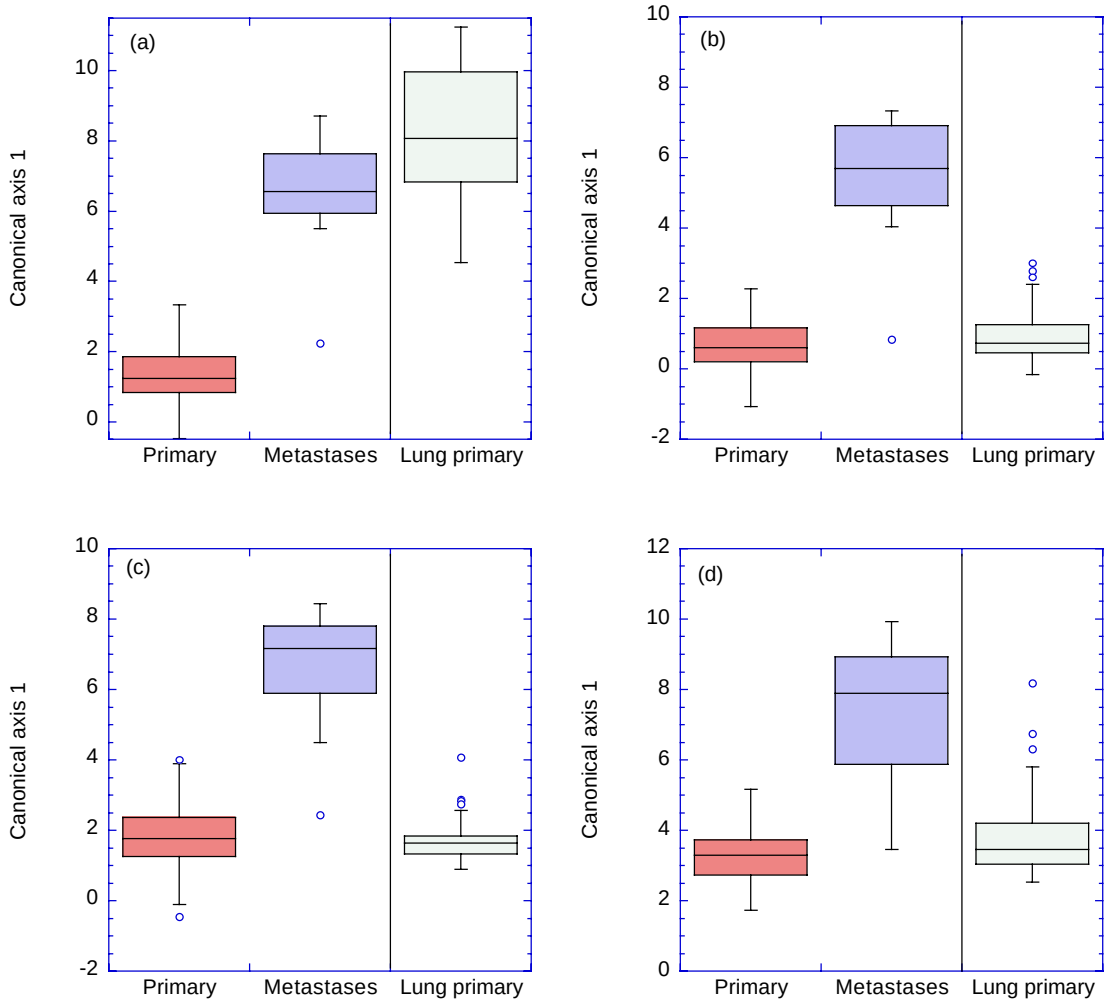


Fig. 6

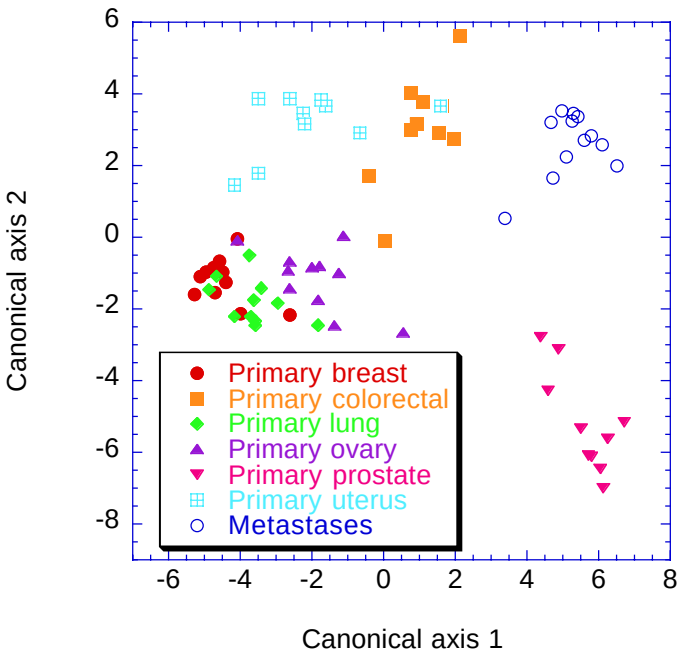

Canonical axis 2
Canonical axis 1
6
4
2
0
-2
-4
-6
-8
-6
-4
-2
0
2
4
6
8
Primary breast
Primary colorectal
Primary lung
Primary ovary
Primary prostate
Primary uterus
Metastases

Fig. 7

UCRL-JRNL-205176

Supplement to Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The General Case for Multiple Groups

J.R. Kercher , J.N. Quong, K.J. Wu, A.A. Quong

Lawrence Livermore National Laboratory, Livermore, California 94551

This article was prepared for journal submission.

July 2004

DISCLAIMER

Supplement to Variable Selection in Canonical Analysis of Gene- and Protein-Expression Data: The General Case for Multiple Groups

J. R.Kercher[1], J. N. Quong[1,2], K. J. Wu[1], A. A. Quong[1,2]

[1]Lawrence Livermore National Laboratory

P.O. Box 808

Livermore, California 94551 USA

[2]Georgetown University, Lombardi Cancer Center, 3970 Reservoir Road, NW,

Washington, DC 20057 USA

{NOTE TO THE EDITOR: We understand that if the main paper is accepted and after the contents are agreeable, this supplement report will be placed online.}

PART A. METHODS OF ANALYSIS

Section I. Notation and Mathematical equivalence of MPGI CA and Subspace CA for two groups.

Notation. We briefly introduce the notation needed for the subspace method. Let the original data be represented by the matrix **X** whose elements x_{ij} are the value of the variable *i* of observation or sample *j*. The index *i* ranges from 1 to *p*, and the index *j* ranges from 1 to *N*. Let $\mathbf{x}_s^j$ denote the *j*th column of **X**. The original data has been classified into *h* groups of observations. Denote the set of indices of the observations of group *k* by J_k where *k*=1,...,*h*. The number of observations in the *k*th group is n_k. That is, $N = \sum_{k=1}^{h} n_k$. The mean value of the variable *i* (protein concentration, gene expression ratio, etc) in the *k*th group is $\bar{x}_{i(k)} = (1/n_k)\sum_{j \in J_k} x_{ij}$ and the grand mean for the entire data set for the *i*th variable is given by $\bar{x}_i = (1/N)\sum_{j=1}^{N} x_{ij}$. The deviation of the *i*th variable of the *j*th observation from the grand mean is $t_{ij} = x_{ij} - \bar{x}_i$. Define the matrices $\mathbf{X_B}$ and $\mathbf{X_W}$ by $(\mathbf{X_B})_{ij} = \sqrt{n_k}\left(\bar{x}_{i(k)} - \bar{x}_i\right)$, and $(\mathbf{X_W})_{ij} = \left(x_{ij} - \bar{x}_{i(k)}\right)$ where $j \in J_k$. Then the within-group sum of squares and cross products (SSCP) matrix is $\mathbf{W} = \mathbf{X_W}\mathbf{X_W^T}$ and the between-group SSCP is given by $\mathbf{B} = \mathbf{X_B}\mathbf{X_B^T}$ where superscript **T** indicates the transpose. Note that the rank of $\mathbf{X_W}$ is at most (*N*–*h*) and therefore the rank of **W** is at most (*N*–*h*).

Singular Value Decomposition of **W**. Because **W** is symmetric, the SVD is given by $\mathbf{V}^{\mathrm{T}}\mathbf{W}\mathbf{V} = \mathbf{\Delta}$; the matrix **V** is nonsingular and orthogonal; and **Δ** is a diagonal matrix with the singular values δ_i on the diagonal. The columns of the *p*×*p* matrix **V** is denoted by $\mathbf{v}^j$. We assign the order of the eigenvalues such that $\delta_i \geq \delta_j$ for *i*<*j*. Note that the SVDs for **W** is equivalent to the eigenvalue equation $\mathbf{W}\mathbf{V} = \mathbf{V}\Delta$. Typically, the rank (number *r* of non-zero eigenvalues) of **W** is almost always (*N*–*h*), which is the number of degrees of freedom of **W**, i.e., δ_i>0 for $i \leq r$. All the other eigenvalues are zero, i.e, $\delta_j = 0$ for *j*>*r*. Define the *r*×*r* diagonal matrix with diagonal elements δ_i, $i \leq r$, as $\mathbf{\Delta_W}$. Define $\mathbf{R_W} = \mathbf{\Delta_W^{-1}}$. Define the *p*×*p* matrix **R** such that $(\mathbf{R})_{ij} = (\mathbf{R_W})_{ij}$ for $i \leq r$, $j \leq r$; $(\mathbf{R})_{ij} = 0$ otherwise. The rank of **B** is almost always the number of degrees of freedom (*h*–1) of **B**.

Equivalence of the MPGI CA and the Subspace CA in the case of two groups. For the case of two groups we demonstrate that the subspace CA is equivalent to the the Moore-Penrose generalized inverse canonical analysis, which was proposed by Kercher et al (2004a). To solve the canonical equation for two groups

$$\mathbf{Be} - \lambda\mathbf{We} = \mathbf{0} \qquad (1)$$

for λ and **e**, where **B** is the between-group sum-of-squares-and-cross-products (SSCP) matrix and **W** is the within-group SSCP, we first note only one eigenvalue ξ_1 of **B** is non-zero whose eigenvector we denote $\mathbf{s}^1$. Note that **B** and **W** are both $p \times p$ matrices where p is the number of variables measured at each sample; there are N samples and h groups. For any nontrivial vector **e**, which is not orthogonal to the eigenvector $\mathbf{s}^1$, eq. 1 can be transformed to

$$\mathbf{We}' = \mathbf{s}^1 \tag{2}$$

where constants λ, ξ_1, and a_1, the coefficient of **e** for $\mathbf{s}^1$ in the expansion of **e** in the eigenvectors of **B**, have been absorbed into the vector $\mathbf{e}'$. The least-squares solution to eq. 2 is

$$\mathbf{e}' = \mathbf{W}^+\mathbf{s}^1 \tag{3}$$

where $\mathbf{W}^+$ is the Moore-Penrose generalized inverse (MPGI). The singular value decomposition of the symmetric matrix **W** (or spectral decomposition) is given by

$$\mathbf{V}^\mathbf{T}\mathbf{WV} = \Delta \tag{4}$$

where the superscript **T** indicates the transpose, **V** is an orthogonal matrix and $\mathbf{\Delta}$ is a diagonal matrix with $r=N-h$ nonzero entries, $\Delta_{ii}>0$ for $i=1,\ldots, r$. All other entries of $\mathbf{\Delta}$ are zero. Denote the upper left $r \times r$ submatrix of $\mathbf{\Delta}$ as $\mathbf{\Delta_W}$. Define $\mathbf{R_W} = \Delta_\mathbf{W}^{-1}$ and $\mathbf{R} = \begin{pmatrix} \mathbf{R_W} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. Denote the first r columns of **V** as the rectangular matrix $\mathbf{V_r}$ and the remaining $p-r$ columns as $\mathbf{V_I}$. Now we may write $\mathbf{W}^+ = \mathbf{VRV}^\mathbf{T}$ (Schott 1997, Thm 5.7). So eq. 3 becomes

$$\mathbf{e}' = \mathbf{V_r}\mathbf{R_W}\mathbf{V_r^T}\mathbf{s}^1 \tag{5}$$

Now consider the subspace method as described in Kercher et al. (2004b). Manipulate eq. 1 to the form

$$\mathbf{V}^\mathbf{T}\mathbf{BVV}^\mathbf{T}\mathbf{e} - \lambda\mathbf{V}^\mathbf{T}\mathbf{WVV}^\mathbf{T}\mathbf{e} = 0 \tag{6}$$

and define

$$\mathbf{f} = \mathbf{V}^\mathbf{T}\mathbf{e}, \tag{7}$$

then eq. 6 becomes

$$\mathbf{V}^\mathbf{T}\mathbf{BVf} - \lambda\Delta\mathbf{f} = 0 \tag{8}$$

Let vector $\mathbf{f_r}$ denote the first *r* entries of **f** and the vector $\mathbf{f_I}$ denote the remaining *p*-*r* entries of **f**. So write eq. 8 as two equations

$$\mathbf{V_r^T B V_r f_r} + \mathbf{V_r^T B V_I f_I} - \lambda \Delta_{\mathbf{W}} \mathbf{f_r} = 0 \qquad (9a)$$

and

$$\mathbf{V_I^T B V_r f_r} + \mathbf{V_I^T B V_I f_I} = 0 \qquad (9b)$$

We shall ignore the second equation, eq. 9b. We pick the particular solution, $\mathbf{f_r}=\mathbf{f_W}$ and $\mathbf{f_I}=\mathbf{0}$, where $\mathbf{f_W}$ is the eigenvector solution to

$$\mathbf{V_r^T B V_r f_W} - \lambda \Delta_{\mathbf{W}} \mathbf{f_W} = 0 \qquad (10)$$

It is convenient to rewrite eq. 10 as

$$\mathbf{f_W} = \lambda^{-1} \mathbf{R_W V_r^T B V_r f_W} \qquad (11)$$

Now we can write the vector **e** (approximate) solution to eq. 1 using eq. 7 and eq. 11 as

$$\mathbf{e} = \mathbf{V_r f_W} = \lambda^{-1} \mathbf{V_r R_W V_r^T B V_r f_W} \qquad (12)$$

Now in the case of two groups as stated above, **B** transforms all nontrivial vectors, which are not orthogonal to $\mathbf{s}^1$, onto $\mathbf{s}^1$. Thus up to a constant of proportionality, which we absorb into **e**, eq. 12 becomes

$$\mathbf{e''} = \mathbf{V_r R_W V_r^T s}^1 \qquad (13)$$

Note that eq. 5 and eq. 13 are identical. This completes the demonstration that the general subspace canonical analysis applied to two groups and the MPGI canonical analysis for two groups are equivalent.

Section II. The nature of the subspace solutions to the canonical equation.

In this section we shall prove two closely related results. The first result is that the subspace method gives an exact solution if the range of **W** and the range of **B** are identical. The second result is that the if the residual vector of the subspace solution to the canonical equation is nonzero, then it lies in the null space of **W**

Theorem 1. If we define $\mathbf{f_W}$ as the solution to eq. 10 and $\mathbf{f_I}=\mathbf{0}$ so that

$$\mathbf{f} = \begin{pmatrix} \mathbf{f_r} \\ \mathbf{f_I} \end{pmatrix} = \begin{pmatrix} \mathbf{f_W} \\ \mathbf{0} \end{pmatrix} \qquad (14)$$

and **e** is given by eq. 7, then

a) $$\left\| \mathbf{Be} - \lambda \mathbf{We} \right\|^2 = \left\| \left(\mathbf{I} - \mathbf{WW}^{+} \right) \mathbf{Be} \right\|^2 \qquad (15)$$

and

b) if all the eigenvectors of **B** with nonzero eigenvalues belong to the range of **W**, then **f** and **e** as defined give exact solutions to eq. 1. Equivalently if the range of **B** is contained in the range of **W**, then the subspace method gives an exact solution to eq. 1.

Proof. Consider the squared residuals of eq. 1 given by the Euclidean vector norm

$$\begin{aligned}\left\|\mathbf{Be}-\lambda\mathbf{We}\right\|^2 &= \left\|\mathbf{Be}-\mathbf{WW}^+\mathbf{Be}+\mathbf{WW}^+\mathbf{Be}-\lambda\mathbf{We}\right\|^2 \\ &= \left\|\mathbf{Be}-\mathbf{WW}^+\mathbf{Be}\right\|^2+\left\|\mathbf{WW}^+\mathbf{Be}-\lambda\mathbf{WW}^+\mathbf{We}\right\|^2 \\ &= \left\|\left(\mathbf{I}-\mathbf{WW}^+\right)\mathbf{Be}\right\|^2+\left\|\mathbf{WW}^+\left(\mathbf{Be}-\lambda\mathbf{We}\right)\right\|^2\end{aligned} \tag{16}$$

Now consider the second term in eq. 16. Using eq. 4 and eq. 7 we rewrite the second term as

$$\begin{aligned}\left\|\mathbf{WW}^+\left(\mathbf{Be}-\lambda\mathbf{We}\right)\right\|^2 &= \left\|\mathbf{V}\Delta\mathbf{R}\mathbf{V}^{\mathbf{T}}\mathbf{BV}\mathbf{V}^{\mathbf{T}}\mathbf{e}-\lambda\mathbf{V}\Delta\mathbf{R}\mathbf{V}^{\mathbf{T}}\mathbf{V}\Delta\mathbf{V}^{\mathbf{T}}\mathbf{e}\right\|^2 \\ &= \left\|\mathbf{V}\left[\begin{pmatrix}\mathbf{I_r} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^{\mathbf{T}}\mathbf{BV}\begin{pmatrix}\mathbf{f_r}\\ \mathbf{f_I}\end{pmatrix}-\lambda\begin{pmatrix}\mathbf{I_r} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\Delta_W & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{f_r}\\ \mathbf{f_I}\end{pmatrix}\right]\right\|^2\end{aligned} \tag{17}$$

where $\mathbf{I_r}$ is the $r\times r$ identity matrix. Now we use the subspace solution defined by eq. 14 and find

$$\begin{aligned}\left\|\mathbf{WW}^+\left(\mathbf{Be_s}-\lambda\mathbf{We_s}\right)\right\|^2 &= \left\|\mathbf{V}\left[\mathbf{V_r^T}\mathbf{BV_r}\mathbf{f_W}-\lambda\Delta_{\mathbf{W}}\mathbf{f_W}\right]\right\|^2 \\ &= \left\|\mathbf{V}\cdot\mathbf{0}\right\|^2=0\end{aligned} \tag{18}$$

where $\mathbf{e_s}$ denotes the subspace solution. Thus we find

$$\left\|\mathbf{Be_s}-\lambda\mathbf{We_s}\right\|^2=\left\|\left(\mathbf{I}-\mathbf{WW}^+\right)\mathbf{Be_s}\right\|^2 \tag{19}$$

Now the SVD of **B** is given by $\mathbf{S^TBS}=\Xi$ where the columns of **S**, which are the eigenvectors of **B**, are denoted $\mathbf{s}^i$ and the eigenvalues ξ_i are the diagonal elements of $\mathbf{\Xi}$. Only h–1 eigenvalues are nonzero. The SVD of **B** is equivalent to the expansion

$$\mathbf{B}=\sum_{i=1}^{h-1}\xi_i\mathbf{s}^i\mathbf{s}^{i\mathbf{T}} \tag{19}$$

Now note that $\mathbf{WW}^+$ is the projection matrix onto the range of **W** by the Moore definition of the Moore-Penrose generalized inverse. Thus if all the eigenvectors of **B** belong to the range of **W**, then

$$\mathbf{WW}^+\mathbf{s}^i=\mathbf{s}^i. \tag{20}$$

for all *i*. Thus $\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{+}\right)\mathbf{B}=\mathbf{0}$, and we find that $\left\|\mathbf{B}\mathbf{e}-\lambda\mathbf{W}\mathbf{e}\right\|^{2}=0$. Thus, we find that the canonical equation eq. 1 is solved exactly if all the eigenvectors of **B** belong to the range of **W**.

Theorem 2. If we define $\mathbf{f}_{\mathbf{W}}$ as the solution to eq. 10 and $\mathbf{f}_{\mathbf{I}}=\mathbf{0}$ so that

$$\mathbf{f}=\begin{pmatrix}\mathbf{f}_{\mathbf{r}}\\ \mathbf{f}_{\mathbf{I}}\end{pmatrix}=\begin{pmatrix}\mathbf{f}_{\mathbf{W}}\\ \mathbf{0}\end{pmatrix}$$

and **e** is given by eq. 7, then the residuals of the canonical equation eq. 1 as given by the subspace solution (eq. 14) are either 0 or lie in the null space of **W**.

Proof. Eq. 18 shows that the residuals have no components in the range of **W**. Eq. 19 can be rewritten

$$\left\|\mathbf{B}\mathbf{e}_{\mathbf{s}}-\lambda\mathbf{W}\mathbf{e}_{\mathbf{s}}\right\|^{2}=\left\|\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{+}\right)\mathbf{B}\mathbf{e}_{\mathbf{s}}\right\|^{2}=\left\|\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{+}\right)\left(\mathbf{B}\mathbf{e}_{\mathbf{s}}-\lambda\mathbf{W}\mathbf{e}_{\mathbf{s}}\right)\right\|^{2}$$

Thus, the length of the residual of the canonical equation is identical to the length of the component of the residual in the null space of **W**.

*Properties of Moore-Penrose generalized inverse***.** In the main paper (Kercher et al. 2004b) and this supplement, we use properties of the MPGI found in standard texts; for example, see Schott (1997, Chap. 5) or Campbell and Meyer (1979, Chap. 1 through 3).

Recasting the SCA for computation. In the standard CA, the canonical equation is $\mathbf{X}_{\mathbf{B}}\mathbf{X}_{\mathbf{B}}^{\mathbf{T}}\mathbf{e}=\lambda\mathbf{X}_{\mathbf{W}}\mathbf{X}_{\mathbf{W}}^{\mathbf{T}}\mathbf{e}$. See Suppl. Sec. I for definitions of $\mathbf{X}_{\mathbf{B}}$ and $\mathbf{X}_{\mathbf{W}}$. When **W** is singular, the canonical equation in the subspace method is $\mathbf{V}_{\mathbf{r}}^{\mathbf{T}}\mathbf{X}_{\mathbf{B}}\mathbf{X}_{\mathbf{B}}^{\mathbf{T}}\mathbf{V}_{\mathbf{r}}\mathbf{f}_{\mathbf{W}}=\lambda\boldsymbol{\Delta}_{\mathbf{W}}\mathbf{f}_{\mathbf{W}}$. Denote the linear combination of variable measurements as the $r\times N$ matrix **U**, i.e., $\mathbf{U}=\mathbf{V}_{\mathbf{r}}^{\mathbf{T}}\mathbf{X}$, $\mathbf{U}_{\mathbf{B}}=\mathbf{V}_{\mathbf{r}}^{\mathbf{T}}\mathbf{X}_{\mathbf{B}}$, $\mathbf{U}_{\mathbf{W}}=\mathbf{V}_{\mathbf{r}}^{\mathbf{T}}\mathbf{X}_{\mathbf{W}}$. The *j*th *r*-dimensional column $\mathbf{u}^{j}$ of **U** is given by $\mathbf{u}^{j}=\mathbf{V}_{\mathbf{r}}^{\mathbf{T}}\mathbf{x}^{j}$. Then the subspace canonical equation (Kercher et al. 2004b, eq. 7) is given by

$$\mathbf{U}_{\mathbf{B}}\mathbf{U}_{\mathbf{B}}^{\mathbf{T}}\mathbf{f}_{\mathbf{W}}=\lambda\boldsymbol{\Delta}_{\mathbf{W}}\mathbf{f}_{\mathbf{W}}. \qquad (21)$$

We note that $\mathbf{V}_{\mathbf{r}}^{\mathbf{T}}$ transforms a *p*-vector measurement (column of **X**) into an *r*-vector pseudo-measurement **u**, a column of **U**, each of whose *r* components is a linear combination of the *p* measurements in the corresponding column of **X**.

Section III. Inference Tests Used in the Canonical Analyses Algorithms

Because each column of **X** is an observation and, by assumption, independent of the other observations (columns) of **X**, then each column of **U** is independent of the other columns of **U**. The new model for the **U** measurements is given by $\mathbf{u}=\mathbf{V}_{\mathbf{r}}^{\mathbf{T}}\mu+\mathbf{V}_{\mathbf{r}}^{\mathbf{T}}\tau^{k}+\mathbf{V}_{\mathbf{r}}^{\mathbf{T}}\varepsilon=\mu'+\tau'^{k}+\varepsilon'$ where ε' is a random variable with the property

$\varepsilon' \sim N_p(\mathbf{0}, \Sigma_\mathbf{r})$. We see that

$$(\Sigma_\mathbf{r})_{ij} = Var(\varepsilon'_i \varepsilon'_j) = E(\varepsilon'_i \varepsilon'_j) - E(\varepsilon'_i)E(\varepsilon'_j) = E(\mathbf{v}^{i\mathrm{T}}\varepsilon \cdot \mathbf{v}^{j\mathrm{T}}\varepsilon) - E(\mathbf{v}^{i\mathrm{T}}\varepsilon)E(\mathbf{v}^{j\mathrm{T}}\varepsilon)$$

$$= \sum_{l=1}^{p} v_l^i \sum_{m=1}^{p} v_m^j \left[E(\varepsilon_l \varepsilon_m) - E(\varepsilon_l)E(\varepsilon_m)\right] = \left(\mathbf{V}_\mathbf{r}^\mathbf{T} \Sigma \mathbf{V}_\mathbf{r}\right)_{ij} \quad ,$$

which applies to all instances of **u** (columns of **U**). Although the **U**-variance matrix $\mathbf{\Sigma}_r$ is different from the **X**-variance matrix $\mathbf{\Sigma}$, all the draws of the random variable ε' are governed by the same variance matrix. Because of the independence and normality properties of **u**, which are based on the independence and sums of normal distributions of the components of **x**, the whole machinery of canonical analysis and statistical inference may be applied to the subspace canonical equation, eq. 21.

At each step in the backward elimination (BE) algorithm, we perform inference tests and find *P*-values for the significance of collections of eigenvectors $\left(P_{\chi^2}(j)\right)$ (e.g., Mardia et al. 1979 p. 343, Cooley and Lohnes 1971 p. 249); significance of individual variables $\left(P_{Ft}(i)\right)$, i.e., coefficients of eigenvectors (Rao 1970, Hawkins 1976, McHenry 1978); pairwise Scheffe comparisons, which are *post hoc* on the variables, for differences of the means of all group pairs (Harris 2001 p. 222); and significance of the Wilks ratio $\left(P_{FW}(i)\right)$ (Rao 1965 p. 471). In the latter test we only make use of rankings based on this probability. We also find confidence intervals for the group means for the CA at each step (e.g., Mardia et al. 1979 p. 345, Seal 1964 p. 137).

Section IV. Implementation of the Subspace CA and Backward Elimination with Ranking Criteria

Description of the Subspace Canonical Analysis Algorithm

Following Morrison's (1990 p. 314), Krzanowski's (2000 p. 66), and Mardia et al.'s (1979 p.219) recommendations for principal components analysis, we recommend that the data be used in the native units if all the units are the same for all variables and in the absence of any additional information. Otherwise, use any *a priori* knowledge to set the scales of the variables such that they all have the same univariate variance. If there is no such *a priori* information and the units differ from variable to variable, then we standardize the data such that the univariate within-group variance of each variable is identical and set to unity. In the examples in the main paper (Kercher et al. 2004b), we work in the native units. This recommendation is made because unlike standard canonical analysis, the first stage of the two-stage procedure, i.e., the SCA algorithm, can be sensitive to units.

Subspace canonical analysis procedure. After reading in the data matrix **X**, we form the within-group data matrix $\mathbf{X_W}$. We perform an SVD on $\mathbf{X_W}$ using the routine DSVDC from the Slatec Library Version 4.1. We use DSVDC to calculate only the first *N* columns of **V** and the eigenvalue matrix $\Delta_\mathbf{W}^{1/2}$. The matrix $\mathbf{X_B}$ is formed, and then $\mathbf{U_B}$ is formed from $\mathbf{X_B}$ and $\mathbf{V_r}$. Then matrices in eq. 9 are formed and the Slatec Library

routine RSG is used to solve eq. 9. The vectors $\mathbf{e}^i$ and $\mathbf{f}_{\mathbf{W}}^i$ are then normalized. Each observation is then projected onto each canonical axis, using eq. 1. We then calculate and display the eight criteria for each of the original variables as described above. Correlations for criteria 6 and 7 and variances for the BW ratio are calculated using the Slatec Library Version 4.1 routine DCOVAR. Depending on the criterion chosen by the user, we then sort the variables according to the chosen criterion. We cut off this list at (*N*–*h*) variables.

Backward elimination algorithm.

Begin the algorithm with (*N*–*h*) variables in the current list. Calculate the CA for the current list and calculate all the inference tests. At each step, find the variable in the current list that would change Wilks ratio Λ the least if it were removed from the current list and calculate the *P*-value for the test due to Rao (1970), Hawkins (1976), and McHenry (1978). Remove that variable and repeat the algorithm until all variables are removed.

Filtering Subspace Canonical Analysis Results for Variable Selection in the Backward Elimination Algorithm

For the convenience of the reader, we reproduce the discussions of criteria 1, part of criterion 6, and the BW ratio from the supplement to Kercher et al (2004a).

Criterion 1: "Sensitivity of between-group SSCP. In this criterion we assume that those variables to which the between group variance in canonical space is most sensitive are the most important. Calculate $\Lambda_{orig} = \det\left|\mathbf{W}_{orig}\right| \big/ \det\left|\mathbf{W}_{orig} + \mathbf{B}_{orig}\right|$ where

$\left(\mathbf{W}_{orig}\right)_{lm} = \sum_{s,o=1}^{p}\sum_{j=1}^{N} E_{ls}^T\left(x_{sj} - \bar{x}_{s(k)}\right)E_{mo}^T\left(x_{oj} - \bar{x}_{o(k)}\right)$ and

$\left(\mathbf{B}_{orig}\right)_{lm} = \sum_{s,o=1}^{p}\sum_{k=1}^{h} E_{ls}^T\left(\bar{x}_{s(k)} - \bar{x}_{s}\right)E_{mo}^T\left(\bar{x}_{o(k)} - \bar{x}_{o}\right)$. Then selecting variable *i* set $E'_{il} = 1.1E_{il}$ for all *l* from 1 to (*h*–1) and $E'_{jl} = E_{jl}$ for all *l* and for $j \neq i$. Then define $\Lambda_{new} = \det\left|\mathbf{W}_{orig}\right| \big/ \det\left|\mathbf{W}_{orig} + \mathbf{B}_{new}\right|$ where $\left(\mathbf{B}_{new}\right)_{lm} = \sum_{s,o\neq i}\sum_{k} E'^T_{ls}\left(\bar{x}_{s(k)} - \bar{x}_{s}\right)E'^T_{mo}\left(\bar{x}_{o(k)} - \bar{x}_{o}\right)$.

Define the sensitivity to the variable *i* as

$\Gamma_i = \left[\left(\Lambda_{orig} - \Lambda_{new}\right)\big/\Lambda_{orig}\right] \Bigg/ \left[\left(\sum_l E'_{il} - \sum_l E_{il}\right)\Bigg/\sum_l E_{il}\right] = 10\cdot\left(\Lambda_{orig} - \Lambda_{new}\right)\big/\Lambda_{orig}$. After calculating the sensitivities of all the variables, we sort on the sensitivities and retain only the (*N*–*h*) largest. These variables are used in the backward elimination."

Criterion 2: Eigenvector coefficient size. We use the eigenvector transformation coefficients to rank the variables in this criterion. Note that $\mathbf{e} = \mathbf{V}_\mathbf{r}\mathbf{f}_\mathbf{r}$ and so $e_j = f_1 v_j^1 + \cdots + f_r v_j^r$. We fix the ranking score RS for the *k*th variable to be

$RS_{2,k} = \sum_{j=1}^{h-1}\left[1 - P_{\chi^2}(j)\right]\sum_{i=1}^{r} f_{ij}^2 v_{ki}^2 W_{ii}\big/(N-h)$. We sort on the ranking scores and retain only the variables with (N–h) largest scores for use in the backward elimination.

Criterion 3: Probability of coefficients and probability of Wilks ratio In this criterion, it is assumed that the probability of significance of the variable k is estimated

by $\text{Prob}_{3,k} = 1 - \prod_{i=1}^{r} \left\{ 1 - v_{ki}^2 \left[1 - P_{FW}(i-1) \right] \left[1 - P_{Ft}(i) \right] \right\}$. Here we treat $\left[1 - P_{Ft}(i) \right]$ as a probability that is conditioned on the probability of the preceding collection of (i–1) variables, which is assumed to be given by $\left[1 - P_{FW}(i-1) \right]$. As in criterion 2, it is assumed that v_{ki}^2 gives the probability of the significance of the original kth variable conditioned on the significance of the subspace variable i, over which we take the product.

Criterion 4: Cumulative probability of coefficients. This is similar to criterion 3, except that it is assumed that the probability of the previous set of subspace variables is the cumulative product of the probabilities of individual coefficients. That is,

$$\text{Prob}_{4,k} = 1 - \prod_{i=1}^{r} \left\{ 1 - v_{ki}^2 \prod_{j=1}^{i} \left[1 - P_{Ft}(j) \right] \right\}.$$

Criterion 6: Maximum correlation. "For this criterion" and the next, "we assume that the correlation of the original variable with the significant canonical axes produced by the subspace CA determines its importance. In criterion 6, we set a significance limit of *a*. If the canonical axis has $P_{\chi^2}(i) < \alpha$, then it is included. The correlation of variable k is found for each of these significant axes and the maximum correlation is used. That is, $RS_{6,k} = \max_{\left\{ i \left| P_{\chi^2}(i) < \alpha \right. \right\}} \left\{ corr\left(x_k, \mathbf{e}^{\mathbf{T}i} \mathbf{x} \right) \right\}$ where

$$corr\left(x_k, \mathbf{e}^{\mathbf{T}i} \mathbf{x} \right) = \left[\sum_{j=1}^{N} \left(x_{kj} - \bar{x}_k \right) \sum_{l=1}^{p} E_{il}^T \left(x_{lj} - \bar{x}_l \right) \right] \Bigg/ \sqrt{ \sum_{j=1}^{N} \left(x_{kj} - \bar{x}_k \right)^2 \sum_{m=1}^{N} \left[\sum_{l=1}^{p} E_{il}^T \left(x_{lm} - \bar{x}_l \right) \right]^2 }$$."

Criterion 7: Total correlation. This ranking criterion differes from criterion 6 only in that we use the correlation summed across all the canonical axes weighted by the estimated probability for that axis. That is, $RS_{7,k} = \sum_{i=1}^{h-1} \left[1 - P_{\chi^2}(i) \right] corr\left(x_k, \mathbf{e}^{\mathbf{T}i} \mathbf{x} \right)$. Note that for the case of two groups, criterion 6 and criterion 7 are identical.

BW ratio. "Ramaswamy et al. (2003) used a univariate signal-to-noise ratio $S_x = (\mu_1 - \mu_2) / (\sigma_1 + \sigma_2)$ where the subscripts refer to group number, μ is the group mean, and σ is the standard deviation within the group. For cases in which the number of groups exceed two, we generalize the signal-to-noise ratio to the ratio of the between-group variance to the within group variance for each variable *i*.

$S_i^2 = \left[\sum_{k=1}^{h} n_{ki} \left(x_{i(k)} - \bar{x}_i \right)^2 \right] \Big/ \sum_{j=1}^{N} \left(x_{ij} - \bar{x}_{i(k)} \right)^2 = (\mathbf{B})_{ii} / (\mathbf{W})_{ii}$. This is the BW ratio of Dudoit et al. (2002)."

PART B. RESULTS

Section V. Significance of eliminated variable in TOFS-SIMS protein data

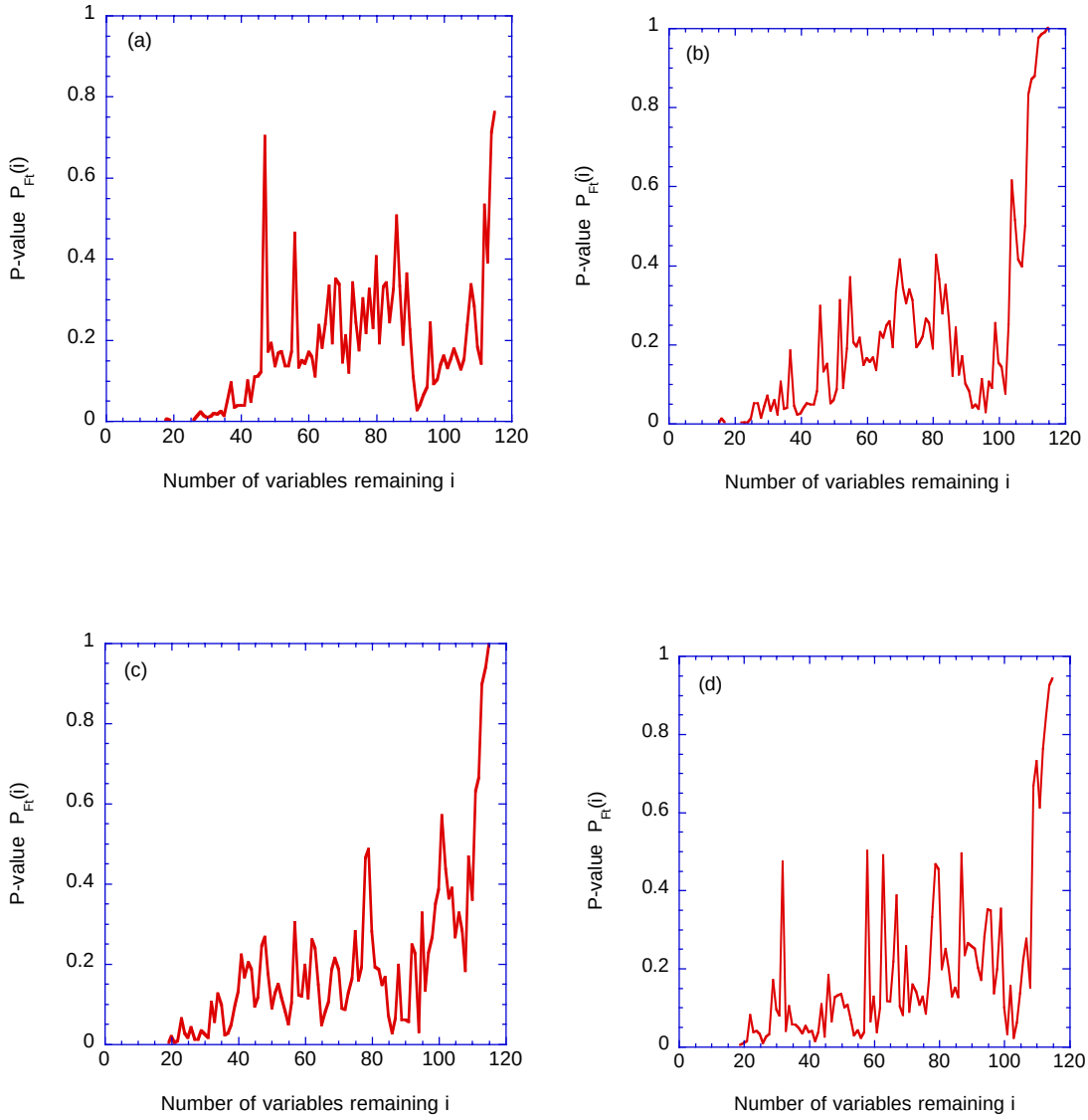


Fig. S.1. See caption on next page.

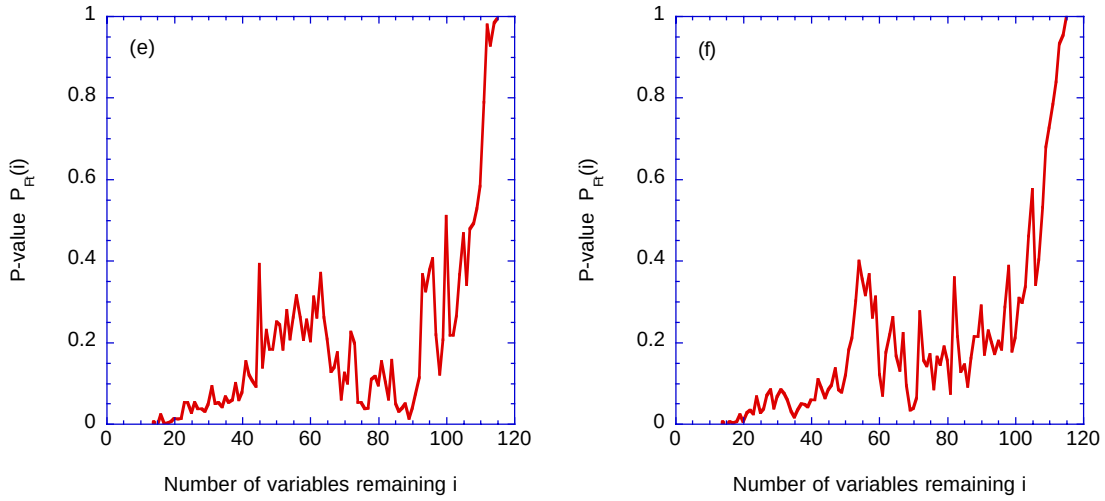


Fig. S1. (continued). Graphs of the *P*-values $P_{Ft}(i)$ in rejecting the null hypothesis that the eliminated protein fragment does not improve the group separation plotted against the number of protein fragments remaining in the set used for the standard canonical analysis. We use probability calculation of Rao (1970), Hawkins (1976), and McHenry (1978). These results are for actual data. Ranking criteria used to provide set of (N–h) variables to backward elimination algorithm: (a) Criterion 2: Eigenvector coefficients, (b) Criterion 3: Probability of coefficients (Significance of eliminated variable) and probability of Wilks ratio, (c) Criterion 4: Cumulative probability of coefficients, (d) Criterion 5: Absolute sensitivity of Wilks ratio, (e) Criterion 6: Maximum correlation of variable with subspace CA axes, and (f) Criterion 7: Total correlation.

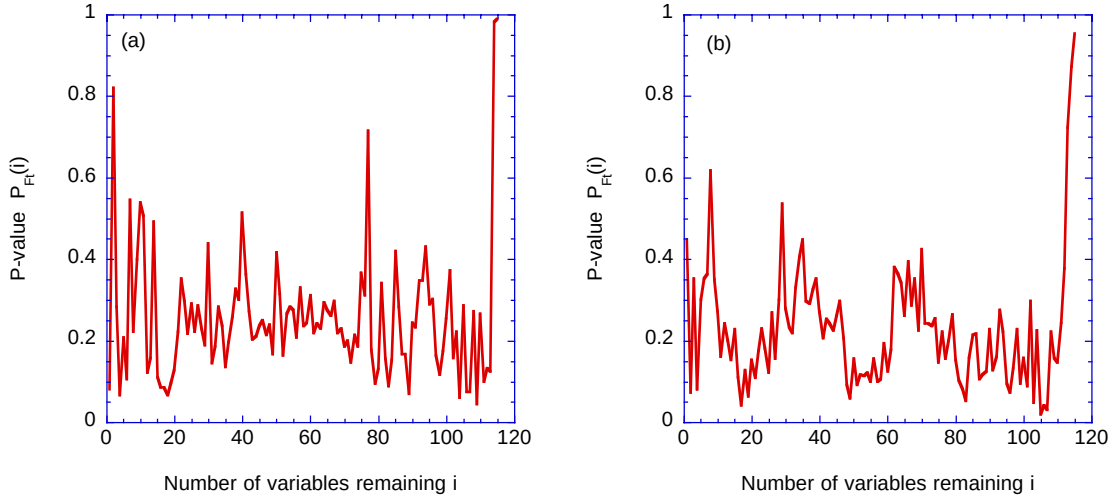


Fig. S2. See caption on next page.

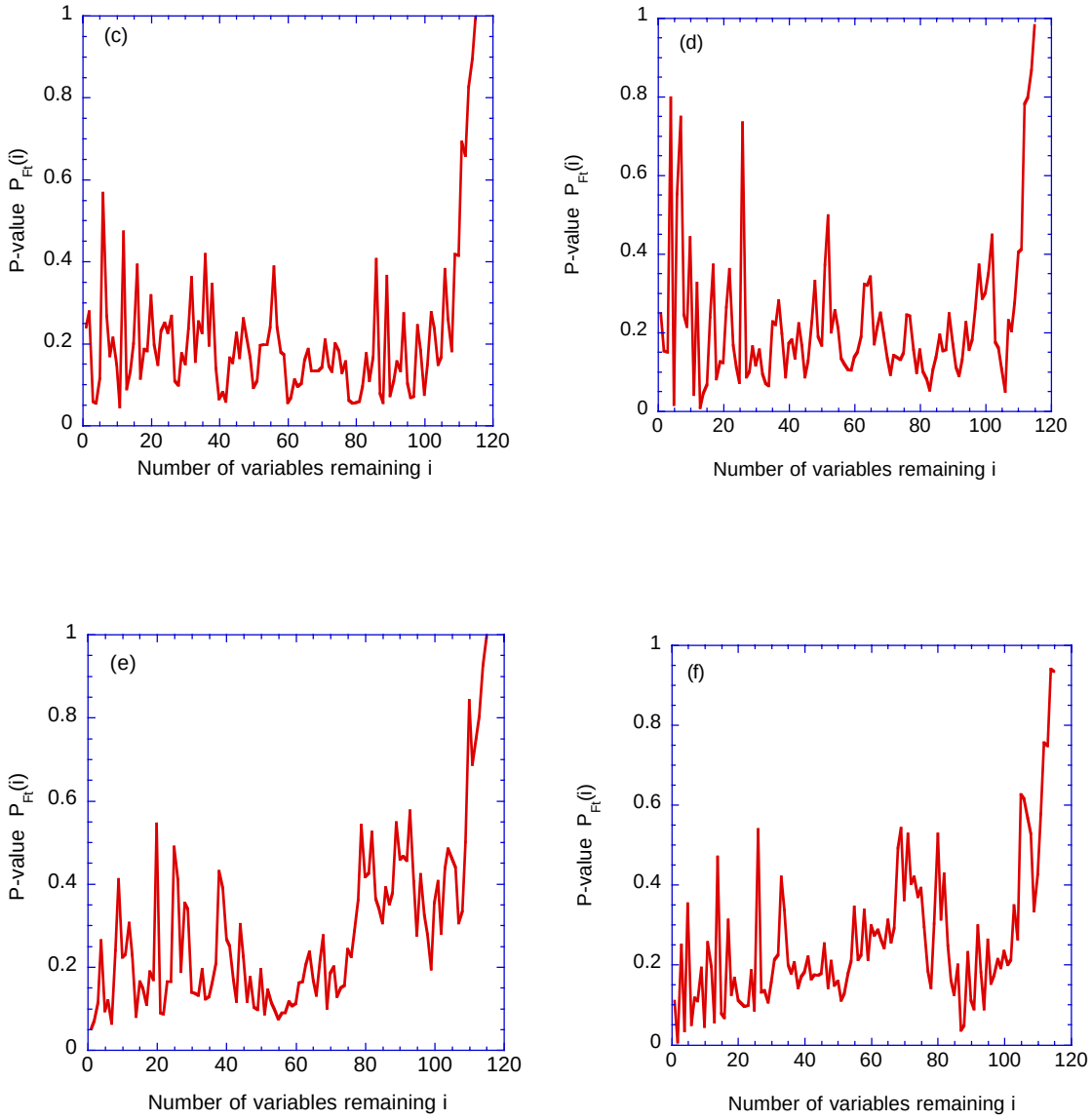


Fig. S2. (continued). Graphs of the *P*-values $P_{Ft}(i)$ in rejecting the null hypothesis that the eliminated protein fragment does not improve the group separation plotted against the number of protein fragments remaining in the set used for the standard canonical analysis. We use probability calculation of Rao (1970), Hawkins (1976), and McHenry (1978). These results are for randomly generated groups. Ranking criteria used to provide set of (*N*–*h*) variables to backward elimination algorithm: (a) Criterion 2: Eigenvector coefficients, (b) Criterion 3: Probability of coefficients (Significance of eliminated variable) and probability of Wilks ratio, (c) Criterion 4: Cumulative probability of coefficients, (d) Criterion 5: Absolute sensitivity of Wilks ratio in canonical space, (e) Criterion 6: Maximum correlation of variable with canonical axes, and (f) Criterion 7: Total correlation.

Section VI. Canonical positions for TOFS-SIMS results for four proteins using criterion 6: Maximum correlation with axes.

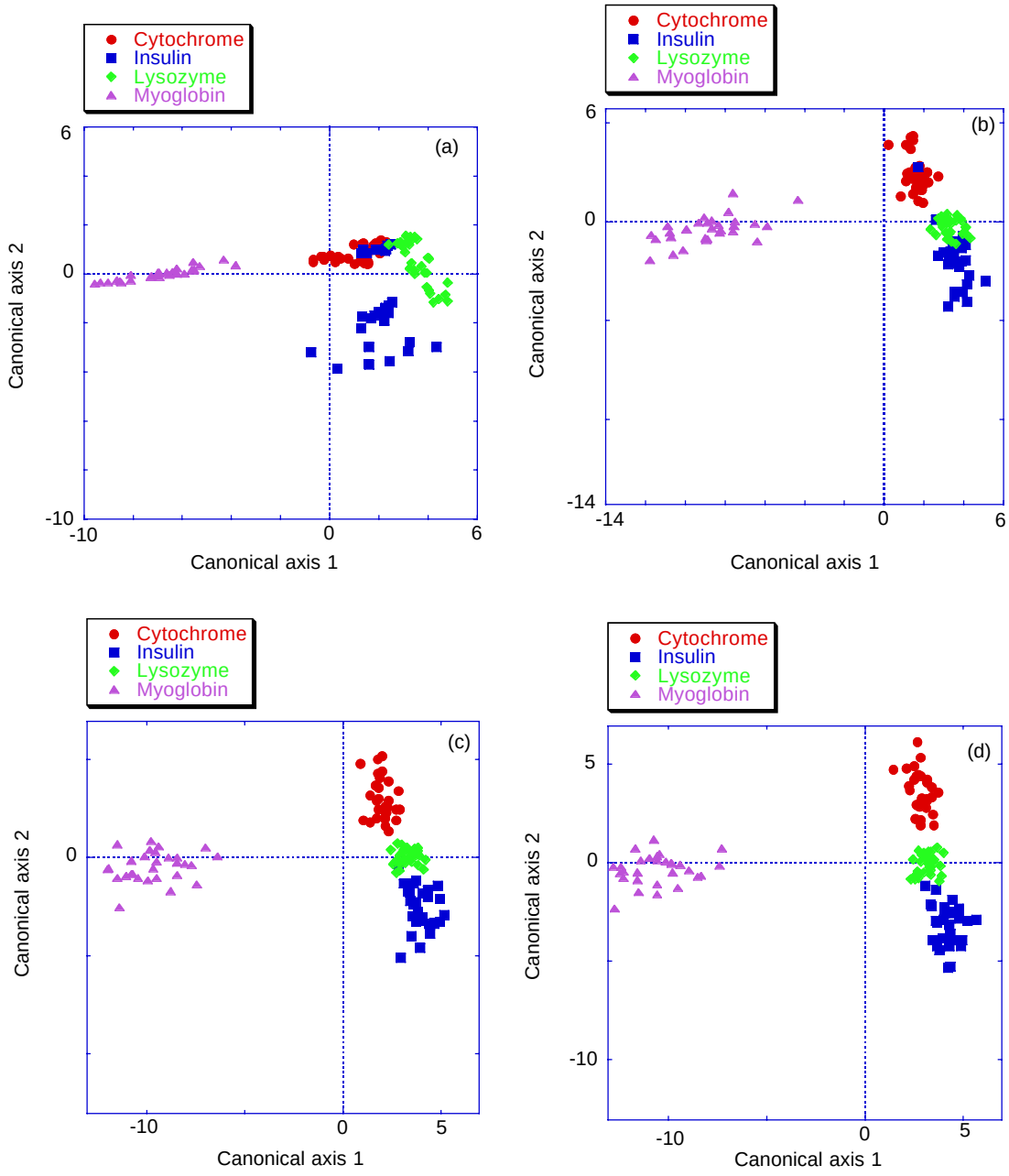


Fig. S.3. Scatterplots of clusters of cytochrome, insulin, lysozyme, and myoglobin fragment data in canonical spaces generated by backward elimination using Criterion 6 (Maximum correlation) to sort fragments for inclusion in the BE procudure. Sensitivities are calculated based on transformations from subspace canonical analysis. Canonical analyses used three, six, nine, and twelve fragments in (a), (b), (c), and (d), respectively.

Section VII. Correlations of variables in canonical signatures with canonical axes.

In Table S.1, we show the twelve fragments defining the twelve-variable CA's for SCA-1-BECA and SCA-6-BECA. Only four fragments (82, 111, 180, and 253 M/Z) are in common. In the Table S.1, we show the sensitivity Γ_i of the Wilks ratio to changes in the coefficients for each variable; a few variables have relatively high sensitivities, while most variables have relatively modest sensitivities. The sensitivity of one of the shared variables (82 M/Z) changes sign from highly negative in the list for criterion 1 to slightly positive in the criterion 6 list. Also, the variable in both lists with the highest positive sensitivity is not one of the four variables held in common. These results might suggest that the space generated by SCA-1-BECA could be different from the space generated by SCA-6-BECA, but a more complete picture is shown in Suppl.B Sec. VIII in which we show the 20 variables with the highest positive and the 20 variables with the highest negative correlations with canonical axes 1 and 2 for the two spaces generated by SCA-1-BECA and SCA-6-BECA. This table shows a very high correspondence between relative positions in the list of correlations with two axes for the two criteria. This high degree of correspondence is true for all fragment variables for these two criteria; note the scatterplots of the ranks of correlation of each fragment with the canonical axes in the figure Suppl.B Sec. VIII. We note that the correlations are more consistent for canonical axis1 than for canonical axis 2. The correlation of the rankings is extraordinarily high for both axes. In Suppl.B Sec. IX, we give the Spearman rank correlations pairwise between all seven criteria and the conventional procedure (BW-BECA) for canonical axis 1 and canonical axis 2, respectively, for the BE CA at twelve variables. For lists 351 elements long, the *P*-values that these rank correlations are due to chance are putatively less than 10^{-12} for all entries in both tables in Suppl.B Sec. IX. We can use the entries in Suppl.B Sec. IX to get an indication for which criteria produce spaces that are similar to the spaces that other criteria produce. Criteria 1 and 2 tend to produce spaces that agree with the other spaces the most, followed by Criteria 4 and 6, and then by Criteria 3, 5, and BW-BECA. Criterion 7 is least likely to produce spaces similar to the other spaces produced. This result is the same for both axes.

The table in Suppl.B Sec. VIII shows that many variables, not used to generate the axes, have higher correlation with the axes than the variables in Suppl.B Sec. VII, which were used to generate the spaces. Most of the 40 variables were not used to generate the CA spaces. A variable does not have to be used in an $\mathbf{e}^i$ of a CA to be correlated with an axis in the new space if there are appropriate correlations with the variables that are used.

In Suppl.B Sec. VIII, we also show the 40 variables with the highest positive or negative correlations averaged across all criteria (columns nine through twelve). These results are for 12-fragment CA's of the BE procedure. We find that fragments 307, 331, 223, 315, and 399 M/Z and fragments 143, 183, 277, 257, and 72 M/Z are the five most highly correlated with canonical axis 1 and canonical axis 2, respectively. That is myoglobin separates from the other protein along an axis highly correlated with 307, 331, 223, 315, and 399 M/Z; while secondarily, all four proteins separate from each other along a gradient correlated with 143, 183, 227, 257, and 72 M/Z.

Table S.1. Twelve protein fragments used in canonical analysis generating Fig. 3d (Criterion 1) and Fig. 4d (Criterion 6). Sensitivity of Wilks ratio (when between-group SSCP changes) to changes in coefficients for fragments (Gi) is in column two. The sensitivity is the percent decrease in the Wilks ratio for a 1 percent increase in the variables coefficients in the **E** matrix. Correlations with each fragment to the first two canonical axes are in columns three and four. Column five gives the order in which the variables were removed from the active list of variables in the BE. Fragment 111 is the last remaining fragment in both cases.

Fragment (M/Z)	Sensitivity	Correlation with canonical axis 1	Correlation with canonical axis 2	Order remaining
		Criterion 1: Sensitivity of between-group SSCP matrix		
64	0.571	0.246	-0.345	7
72	0.463	0.214	-0.546	5
82	-2.808	0.135	-0.237	3
111	1.194	-0.531	-0.290	1
112	0.521	0.263	-0.290	9
136	2.061	0.424	-0.138	6
161	0.006	0.395	-0.357	8
180	0.537	0.329	-0.417	10
183	0.733	0.482	-0.587	4
253	0.169	-0.234	-0.042	12
282	0.226	-0.710	-0.133	11
329	0.716	-0.626	-0.144	2
		Criterion 6: Maximum correlation with canonical axes		
60	0.265	0.316	–0.563	3
82	0.142	0.081	–0.223	4
103	1.672	0.086	–0.469	6
111	0.541	–0.572	–0.239	1
147	-0.515	–0.563	–0.149	8
165	0.211	–0.621	–0.419	7
180	0.173	0.264	–0.424	12
203	0.469	0.323	–0.536	5
208	0.570	–0.589	–0.256	9
253	0.433	–0.274	0.005	10
257	0.369	0.146	0.601	11
328	0.531	–0.643	–0.086	2

Section VIII. Correlations of variables with canonical axes.

Table S.2. The 40 highest correlations of canonical axes 1 and 2 with original variables for the 12-protein backward elimination CA's shown in Fig. 3d (Kercher et al. 2004b) and Fig. S.3d. The last four columns contain the correlations of the fragment variables averaged over all criteria. The fragment designations are in the columns labeled M/Z. The correlations are in the columns labeled by the canonical axis with which the fragment variable was correlated. The average was taken over the 12-protein CA's generated during the BE procedure of the TOFS-SIMS data.

Criterion 1: Sensitivity of Wilks				Criterion 6: Maximum corrrelation				Average over all criteria			
M/Z	Axis 1	M/Z	Axis 2	M/Z	Axis 1	M/Z	Axis 2	M/Z	Axis 1	M/Z	Axis 2
Negative correlations											
307	-0.84	143	-0.60	307	-0.87	143	-0.61	307	-0.83	143	-0.59
331	-0.82	183	-0.59	331	-0.84	183	-0.60	331	-0.80	183	-0.58
223	-0.77	277	-0.58	223	-0.81	277	-0.57	223	-0.76	277	-0.57
315	-0.76	72	-0.55	315	-0.79	60	-0.56	315	-0.76	72	-0.52
399	-0.76	60	-0.54	399	-0.78	201	-0.54	399	-0.74	60	-0.52
398	-0.74	101	-0.53	398	-0.77	178	-0.54	355	-0.72	219	-0.52
355	-0.74	201	-0.53	355	-0.75	203	-0.54	398	-0.72	201	-0.52
330	-0.72	219	-0.52	330	-0.75	202	-0.54	330	-0.70	101	-0.51
281	-0.71	190	-0.52	356	-0.73	219	-0.53	281	-0.70	220	-0.51
282	-0.71	229	-0.52	57	-0.73	101	-0.53	57	-0.70	89	-0.51
57	-0.71	220	-0.52	282	-0.72	72	-0.53	282	-0.69	190	-0.51
356	-0.71	89	-0.51	281	-0.72	89	-0.53	85	-0.69	144	-0.50
85	-0.70	202	-0.50	85	-0.72	220	-0.53	356	-0.68	202	-0.50
358	-0.69	263	-0.50	358	-0.72	144	-0.52	357	-0.68	203	-0.50
357	-0.68	144	-0.50	357	-0.71	190	-0.52	358	-0.67	178	-0.50
371	-0.67	178	-0.50	371	-0.69	179	-0.51	371	-0.67	229	-0.49
383	-0.67	203	-0.50	383	-0.69	229	-0.51	383	-0.65	200	-0.49
388	-0.67	200	-0.50	397	-0.69	200	-0.51	207	-0.65	179	-0.48
207	-0.67	214	-0.49	207	-0.68	227	-0.51	385	-0.65	204	-0.48
385	-0.66	179	-0.49	385	-0.68	181	-0.51	301	-0.65	227	-0.48
Positive correlations											
118	0.43	255	0.06	119	0.37	270	0.11	77	0.43	338	0.07
185	0.43	380	0.08	153	0.37	378	0.11	118	0.43	255	0.07
77	0.43	212	0.09	91	0.37	322	0.12	136	0.43	212	0.07
160	0.43	242	0.09	138	0.37	242	0.13	160	0.43	312	0.08
138	0.43	312	0.10	154	0.38	312	0.13	138	0.43	242	0.09
153	0.43	271	0.12	77	0.38	255	0.13	153	0.43	58	0.11
92	0.44	322	0.13	136	0.38	394	0.15	92	0.43	322	0.12
154	0.44	58	0.14	168	0.38	58	0.17	154	0.44	373	0.14
124	0.44	373	0.16	92	0.39	271	0.17	168	0.44	271	0.15
168	0.44	394	0.17	137	0.39	373	0.19	124	0.44	394	0.16
156	0.45	340	0.18	156	0.40	340	0.22	137	0.45	400	0.19
137	0.45	284	0.22	124	0.40	400	0.23	156	0.45	340	0.20
70	0.45	400	0.22	183	0.41	372	0.25	70	0.46	284	0.20
140	0.47	372	0.26	70	0.42	284	0.27	183	0.47	372	0.22
152	0.48	254	0.39	152	0.42	84	0.45	140	0.48	254	0.40
183	0.48	240	0.43	140	0.42	254	0.46	152	0.48	240	0.41
107	0.49	84	0.45	202	0.43	240	0.47	142	0.49	84	0.43
142	0.49	258	0.46	142	0.44	258	0.51	107	0.49	258	0.45
202	0.51	256	0.53	107	0.44	256	0.58	202	0.50	256	0.52
184	0.52	257	0.54	184	0.48	257	0.60	184	0.53	257	0.53

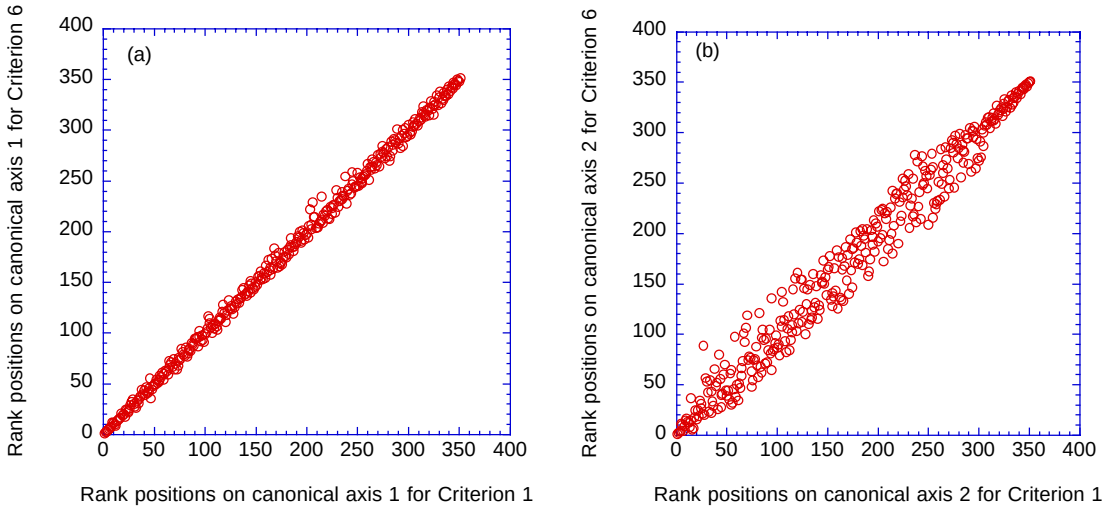


Fig. S.4. Scatterplots of ranks of correlations of 351 fragments with canonical axes comparing spaces determined from using Criterion 6 and Criterion 1 in selection for BE procedure. (a) Ranks found for correlations of fragment variables with canonical axis1; (b) canonical axis 2.

Section IX. Spearman rank correlations pairwise between correlations on canonical axes.

Table S.3. Spearman rank correlation for ranks of correlations of fragments with canonical axis 1. The value at the intersection of a row and a column is the Spearman rank correlation of the ranks of lists of correlations of fragments with the canonical axis produced by the two criteria listed in the row heading and the column heading. *P*-values of these rank correlations are less than 10^{-12} for lists 351 elements long. These axes were generated by canonical analysis at twelve variables from the BE procedure for protein-fragment variables in TOFS-SIMS analysis of four proteins.

Criteria/ Procedure	SCA-2-BECA	SCA-3-BECA	SCA-4-BECA	SCA-5-BECA	SCA-6-BECA	SCA-7-BECA	BW-BECA
SCA-1-BECA	0.9991	0.9985	0.9983	0.9911	0.9988	0.9916	0.9983
SCA-2-BECA		0.9964	0.9993	0.9944	0.9987	0.9895	0.9988
SCA-3-BECA			0.9949	0.9852	0.9971	0.9923	0.9960
SCA-4-BECA				0.9949	0.9977	0.9891	0.9981
SCA-5-BECA					0.9938	0.9726	0.9952
SCA-6-BECA						0.9857	0.9996
SCA-7-BECA							0.9845
Average correlations to other CA's							
SCA-1-BECA	SCA-2-BECA	SCA-3-BECA	SCA-4-BECA	SCA-5-BECA	SCA-6-BECA	SCA-7-BECA	BW-BECA
0.9965	0.9966	0.9943	0.9960	0.9896	0.9959	0.9865	0.9958
Rank order of correlations							
2	1	6	3	7	4	8	5

Table S.4. Spearman rank correlation for ranks of correlations of fragments with canonical axis 2. The value at the intersection of a row and a column is the Spearman rank correlation of the ranks of lists of correlations of fragments with the canonical axis produced by the two criteria listed in the row heading and the column heading. *P*-values of these rank correlations are less than 10^{-12} for lists 351 elements long. These axes were generated by canonical analysis at twelve variables from the BE procedure for protein-fragment variables in TOFS-SIMS analysis of four proteins.

Criteria/ Procedure	SCA-2-BECA	SCA-3-BECA	SCA-4-BECA	SCA-5-BECA	SCA-6-BECA	SCA-7-BECA	BW-BECA
SCA-1-BECA	0.993	0.971	0.976	0.941	0.985	0.662	0.932
SCA-2-BECA		0.952	0.976	0.959	0.991	0.603	0.949
SCA-3-BECA			0.948	0.873	0.938	0.771	0.845
SCA-4-BECA				0.970	0.981	0.570	0.891
SCA-5-BECA					0.975	0.416	0.923
SCA-6-BECA						0.562	0.947
SCA-7-BECA							0.477
Average correlations with other CA's							
SCA-1-BECA	SCA-2-BECA	SCA-3-BECA	SCA-4-BECA	SCA-5-BECA	SCA-6-BECA	SCA-7-BECA	BW-BECA
0.923	0.9178	0.900	0.902	0.865	0.911	0.580	0.852
Rank order of correlations							
1	2	5	4	6	3	8	7

Section X. Comparison of fragment variables selected by different CA's

In Table S.5, we show the fragments used by the CA's in the BE when only 14 fragments remain. At the bottom of the table, we give the number of unique fragments for each criterion. For example, the four M/Z fragments 66, 91, 159, and 170 are unique to Criterion 4. By this measure in the 14-variable case, Criteria 1 and 2 produce lists of fragments that are most like the other lists, followed by criteria 3 and 4, followed by criteria 5 and 6, followed by criterion 7, and BW-BECA.

Table S.5. Fragments selected in the BE at the common number of 14 variables. Results are from BE canonical analyses of four proteins using TOF-SIMS.

Criterion 1: Sens.	Criterion 2: Coeff.	Criterion 3: Prob W and axes	Criterion 4: Cum Prob axes	Criterion 5: Abs. Sens.	Criterion 6: Max. Correl-ation	Criterion 7: Total Correla-tion	BW-BECA (conventional)
64	52	52	60	52	60	57	107
70	57	61	61	57	82	77	111
72	64	70	66	59	103	84	143
82	70	82	69	61	111	87	152
84	71	84	70	69	147	108	164
111	82	96	71	70	164	121	165
112	88	108	72	71	165	126	168
136	95	113	84	84	180	128	183
161	120	120	91	119	203	136	191
180	129	131	111	136	208	183	193
183	136	145	121	142	253	184	215
253	143	147	159	151	257	193	244
282	183	165	170	167	282	204	291
329	207	183	183	207	328	241	316
			Number of unique fragments				
3	3	4	4	5	5	7	8

Section XI. Class prediction in TOF-SIMS protein data

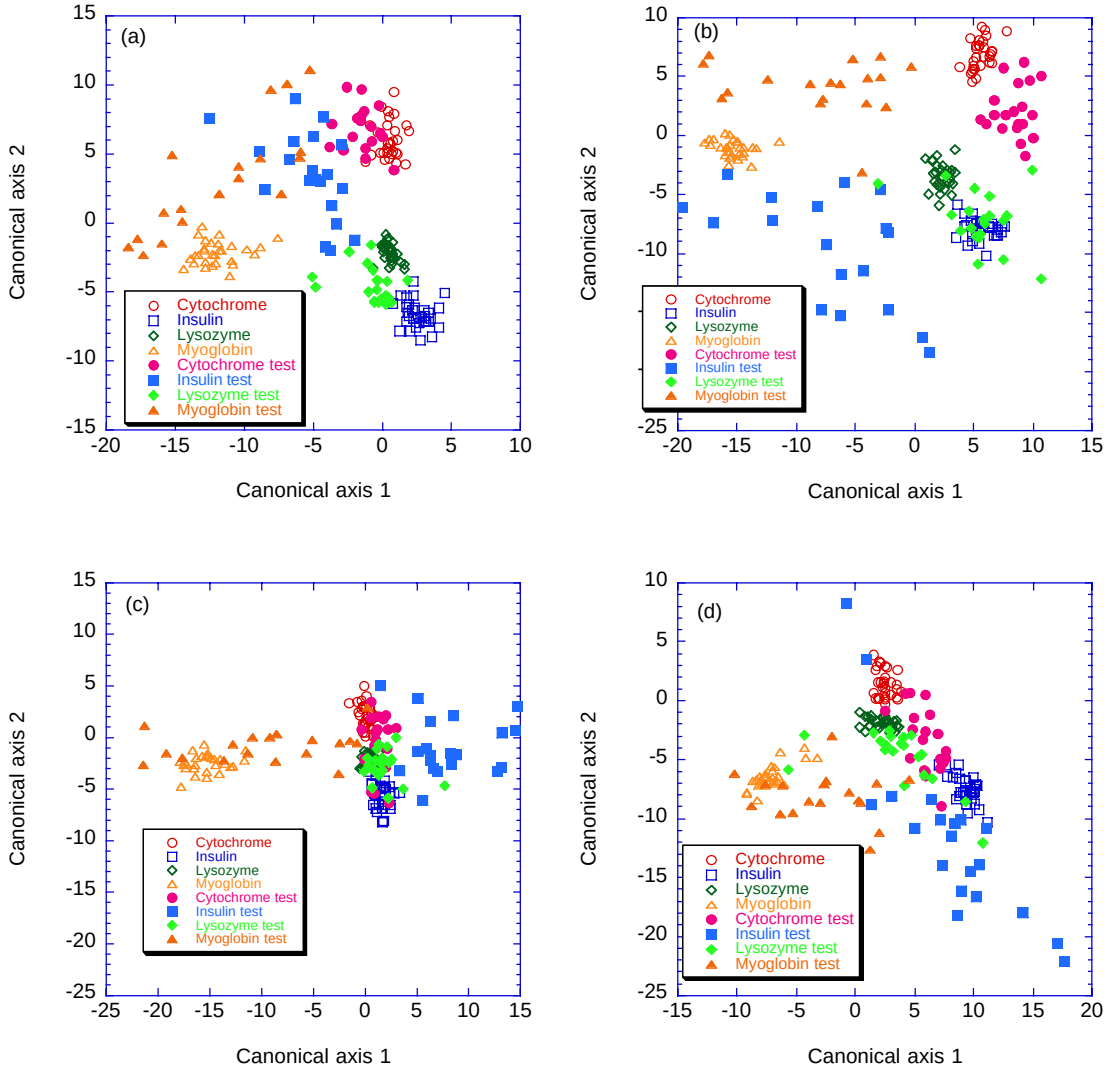


Fig. S.5. (a) SCA-3-BECA 15 fragments, (b) SCA-5-BECA 16 fragments, (c) SCA-6-BECA 12 fragments, and (d) SCA-7-BECA 13 fragments.

Section XII. Class prediction in micro-array tumor data

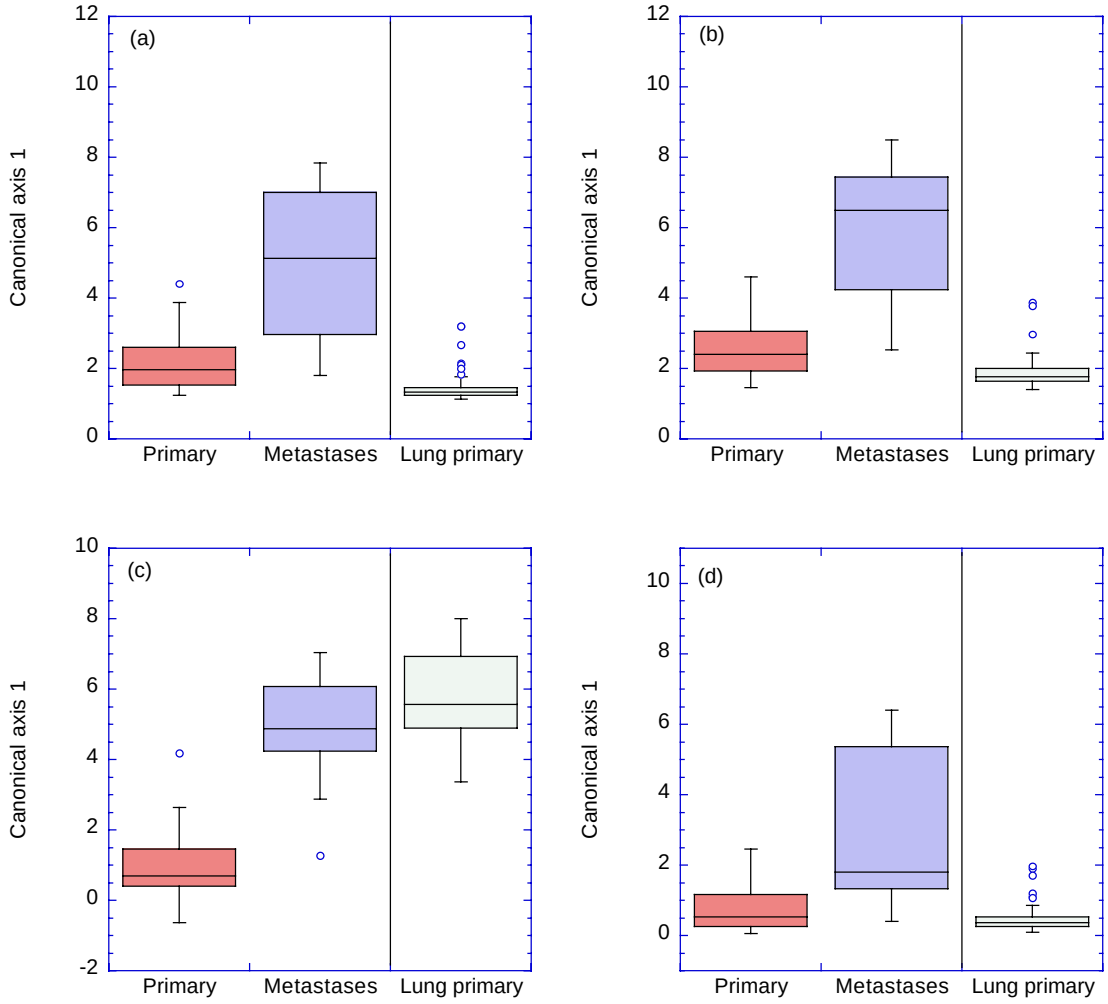


Fig. S.6. Training set consists of primary tumors and metastases shown on left. Test set is shown on right consisting of lung primary tumors. Data from Ramaswamy et al. (2003). (a) SCA-1-BECA Bonferroni-adjusted cutoff at 3 genes. (b) SCA-1-BECA Cutoff at 5 genes. (c) SCA-2-BECA Cutoff at 4 genes. (d) SCA-3-BECA Bonferroni-adjusted cutoff at 2 genes.

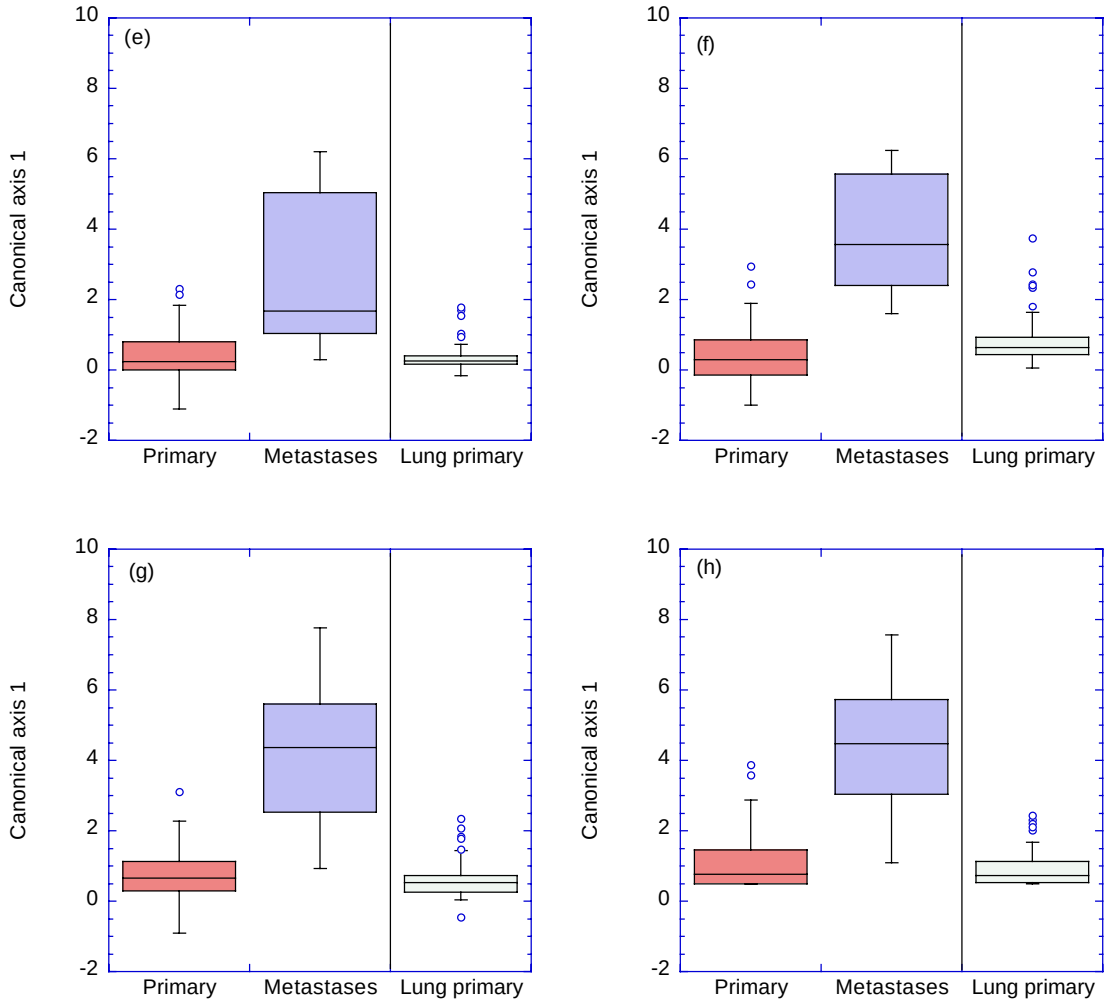


Fig. S.6 (e) SCA-3-BECA Cutoff at 3 genes. (f) SCA-3-BECA 9 genes. (g) SCA-4-BECA Cutoff at 4 genes. (h) SCA-6,7-BECA Bonferroni-adjusted cutoff at 2 genes.

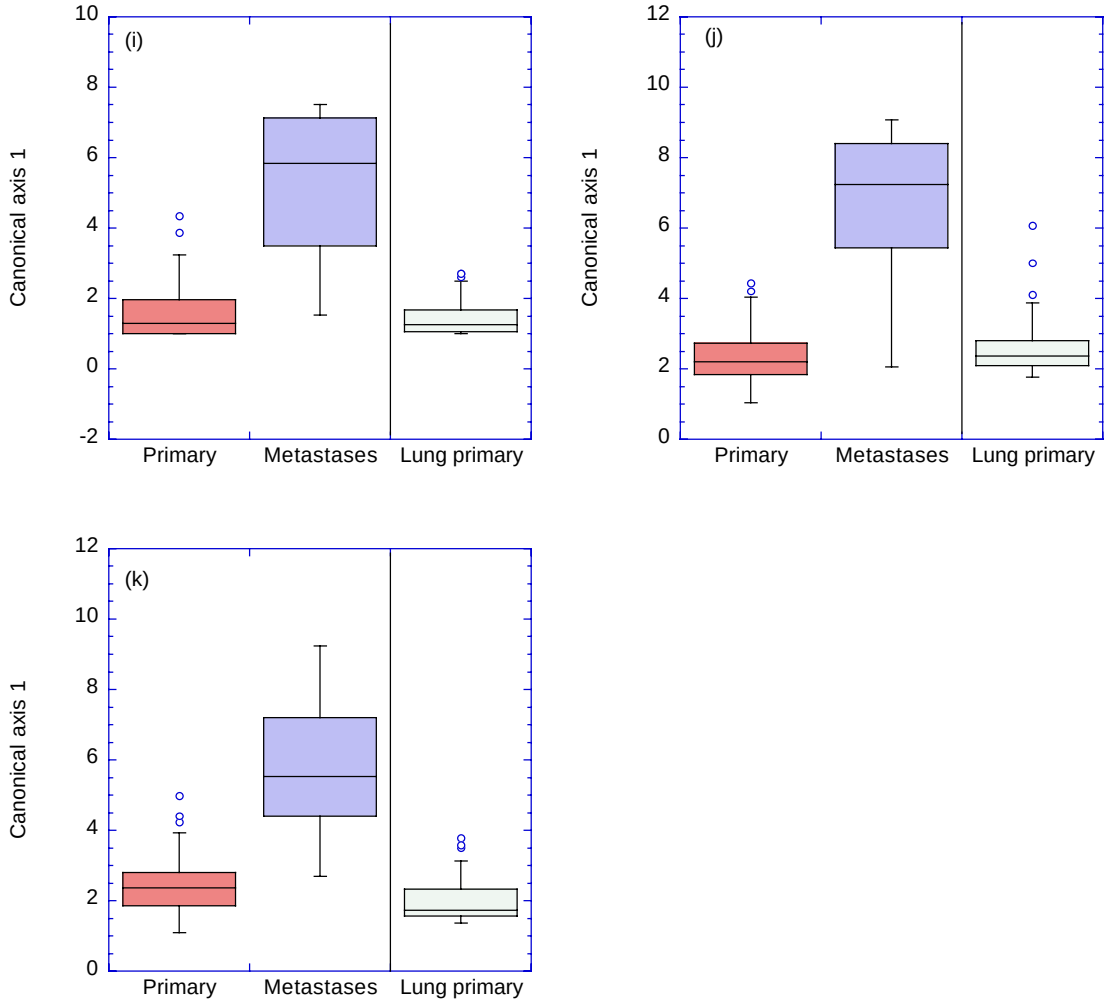


Fig. S.6. (i) SCA-6,7-BECA Cutoff at 3 genes. (j) SCA-6,7-BECA at 8 genes. (k) BW-BECA Cutoff at 4 genes.

Section XIII. Transformation vectors $\mathbf{e}^1$ for primary and metastases data.

Table S.6. Transformation vectors $\mathbf{e}^1$ used to map microarray measurements to canonical space for Table 2 in Kercher et al. (2004b). Data from Ramaswamy et al. (2003).

GenBank ID	$\mathbf{e}^1$ terms Bonf. Cutoff	$\mathbf{e}^1$ terms Cutoff	$\mathbf{e}^1$ terms Lowend	GenBank ID	$\mathbf{e}^1$ terms Bonf. Cutoff	$\mathbf{e}^1$ terms Cutoff	$\mathbf{e}^1$ terms Lowend
SCA-1-BECA				**SCA-5-BECA**			
AA460436	1.344	0.963		X82494	0.756	0.756	1.209
U65410	4.220	4.599		AA195031	0.198	0.198	0.222
Z74615	0.030	0.020		Z14244	0.162	0.162	0.169
K03515		0.052		AA608850			0.159
AA412620		0.371		AA449951			-0.318
SCA-6,7-BECA				AA400410			0.347
X82494	0.916	0.802	0.708	AA100089			-0.226
AA460436	1.515	1.481	1.475	Z74616			-0.046
U65410		2.750	5.652	AA037386			-0.177
AA449951			-0.304	AA412059			0.364
L37747			1.174	X85372			0.393
AA428024			0.187	**BW-BECA (conventional)**			
AA037386			-0.144	S80437	0.075	0.067	0.103
AA486831			0.126	U65410	4.903	4.621	8.642
SCA-3-BECA				K03515	0.081	0.079	0.078
AA609674	0.202	0.308	0.344	D89377		0.749	0.605
S80437	0.073	0.072	0.100	AA009596			0.431
L29433		–0.0494	-0.480	AA093131			-0.237
AA600140		–0.031	-0.078	AA486831			0.200
AA428172			0.090	AA449951			-0.408
AA400410			0.208	L37747			1.521
AA093131			-0.243	J02783			-0.028
U38864			-0.364	AA096094			0.047
K03515			0.050	AA037386			-0.148
SCA-2-BECA				**SCA-4-BECA**			
X82494	0.917	0.865	1.247	X82494	0.912	0.798	1.069
AA460436	1.515	1.293	1.162	AA496788	0.783	0.782	0.680
AA010619		0.773	1.090	U32645		0.840	0.753
M69177		-0.178	-0.222	L39833		–0.994	-1.007
AA496788			0.865	AA412620			0.483
AA491234			-0.201	AA100089			-0.126
X85372			0.465	AA428172			-0.073
AA485358			-0.153	X81900			0.038
AA296994			0.048	AA009596			-0.384
HG4264-HT4534			-0.182	AA598680			0.444
				U23946			-0.433
				U45974			0.119
				X12458			-0.380

Section XIV. Correlations of individual gene expression with canonical axis

Table S.7. Correlation C_i of gene variable *i* with canonical axis for the two-group backward elimination canonical analyses shown in Table 2 (Kercher et al. 2004b). We show the 15 genes with the highest negative correlation.

GenBank ID	C_i	GenBank ID	C_i	GenBank ID	C_i	GenBank ID	C_i
Criterion 1: 5 genes		Criterion 2: 10 genes		Criterion 3: 9 genes		Criterion 4: 13 genes	
AA460436	-0.686	J03464	-0.632	AA400410	-0.673	J03464	-0.660
AA400410	-0.683	X82494	-0.632	AA460436	-0.668	X82494	-0.641
J03464	-0.675	AA460436	-0.606	J03464	-0.648	AA460436	-0.598
AA252812	-0.661	AA400410	-0.573	AA609674	-0.622	AA195031	-0.570
AA449951	-0.621	Z74616	-0.562	AA100089	-0.620	AA400410	-0.568
AA227448	-0.619	AA096094	-0.549	AA093131	-0.610	U75285	-0.544
AA025213	-0.619	L37747	-0.548	AA485358	-0.607	Z74616	-0.537
AA236972	-0.615	U75285	-0.543	AA296994	-0.602	AA025213	-0.537
U75285	-0.611	AA195031	-0.534	U75285	-0.600	AA598680	-0.536
AA195031	-0.610	AA025213	-0.528	AA096094	-0.600	U32645	-0.530
X85372	-0.604	AA608850	-0.526	X81900	-0.598	AA496788	-0.527
AA609674	-0.598	X85372	-0.526	AA236972	-0.591	AA621096	-0.517
AA009596	-0.597	AA010619	-0.519	AA496788	-0.591	AA227448	-0.516
Z74616	-0.596	AA496788	-0.519	S80437	-0.591	HG4264-HT4	-0.515
AA598680	-0.594	AA296994	-0.519	AA195031	-0.585	AA340293	-0.514
Criterion 5: 11 genes		Criterion 6-7: 8 genes		BW-BECA: 12 genes			
X82494	-0.639	X82494	-0.662	J03464	-0.677		
AA195031	-0.615	AA460436	-0.635	AA400410	-0.645		
AA400410	-0.599	J03464	-0.628	AA460436	-0.607		
J03464	-0.593	AA252812	-0.599	U75285	-0.593		
HG4264-HT4534	-0.580	AA400410	-0.597	Z74616	-0.591		
AA460436	-0.550	AA428024	-0.583	X85372	-0.587		
AA025213	-0.545	L37747	-0.555	AA195031	-0.587		
AA096094	-0.538	AA227448	-0.554	AA252812	-0.565		
AA428024	-0.535	Z74616	-0.553	AA025213	-0.565		
X85372	-0.532	AA195031	-0.552	AA093131	-0.561		
Y10807	-0.532	U75285	-0.548	AA496788	-0.555		
HG110-HT110	-0.530	AA025213	-0.540	L37747	-0.554		
AA093131	-0.530	AA412620	-0.536	AA236972	-0.548		
U51166	-0.529	U65410	-0.534	AA485358	-0.544		
U75285	-0.529	AA486831	-0.531	S80437	-0.543		

Table S.8. Correlation C_i of gene variable i with canonical axis for the two-group backward elimination canonical analyses shown in Table 2 (Kercher et al. 2004b). We show the 10 genes with the highest positive correlation.

GenBank ID	C_i	GenBank ID	C_i	GenBank ID	C_i	GenBank ID	C_i
Criterion 1: 5 genes		Criterion 2: 10 genes		Criterion 3: 9 genes		Criterion 4: 13 genes	
AF001548	0.249	M26061	0.293	L29433	0.278	X91103	0.271
L40411	0.250	X91103	0.295	D43968	0.283	D38553	0.275
M27830	0.257	S72043	0.301	M26061	0.284	X14329	0.285
M26061	0.261	X14329	0.309	U45448	0.292	M26061	0.295
M83664	0.282	HG4660-HT5073	0.309	U38864	0.296	Y08639	0.296
D43968	0.290	AA156670	0.315	M27830	0.301	M83664	0.303
U45448	0.294	U45448	0.328	Y08639	0.309	S72043	0.306
S72043	0.302	D43968	0.339	S67156	0.343	U45448	0.324
S67156	0.321	U45974	0.339	X66141	0.365	D43968	0.369
X66141	0.337	X66141	0.371	S72043	0.371	X66141	0.412
Criterion 5: 11 genes		Criterion 6-7: 8 genes		BW-BECA: 12 genes			
D38553	0.302	S79281	0.252	J04970	0.254		
X87767	0.309	M26061	0.255	M83664	0.258		
W52686	0.321	HG4660-HT5073	0.255	U45448	0.267		
X14329	0.324	M10098	0.263	S72043	0.276		
S72043	0.328	U45974	0.263	M63589	0.278		
HG4660-HT5073	0.331	S72043	0.265	M27830	0.283		
U45974	0.336	X94612	0.267	D43968	0.286		
U45448	0.349	S67156	0.272	M26061	0.310		
X66141	0.355	X66141	0.309	X66141	0.341		
D43968	0.399	M27830	0.317	S67156	0.400		

Section XV. Overfit in class prediction as the number of variables increases

Effect of increasing variable count on predictability . In Fig. S.7 we show the boxplots and predictions for criterion 6-7 for 34 genes. The predicted location for the lung primary tumors has moved substantially away from the location of the 128-gene-data-set primary tumors, though it does remain closer to the original primary tumor group than to the metastases group. In Fig. S.7 we show the boxplots and predictions for criterion 1 using 46 genes. Here the original primary group and the metastases are separated, but the predicted location of the lung primary tumors is far from both of the original primary tumor group and the metastases group, and is even closer to the metastases group than the original primary group.

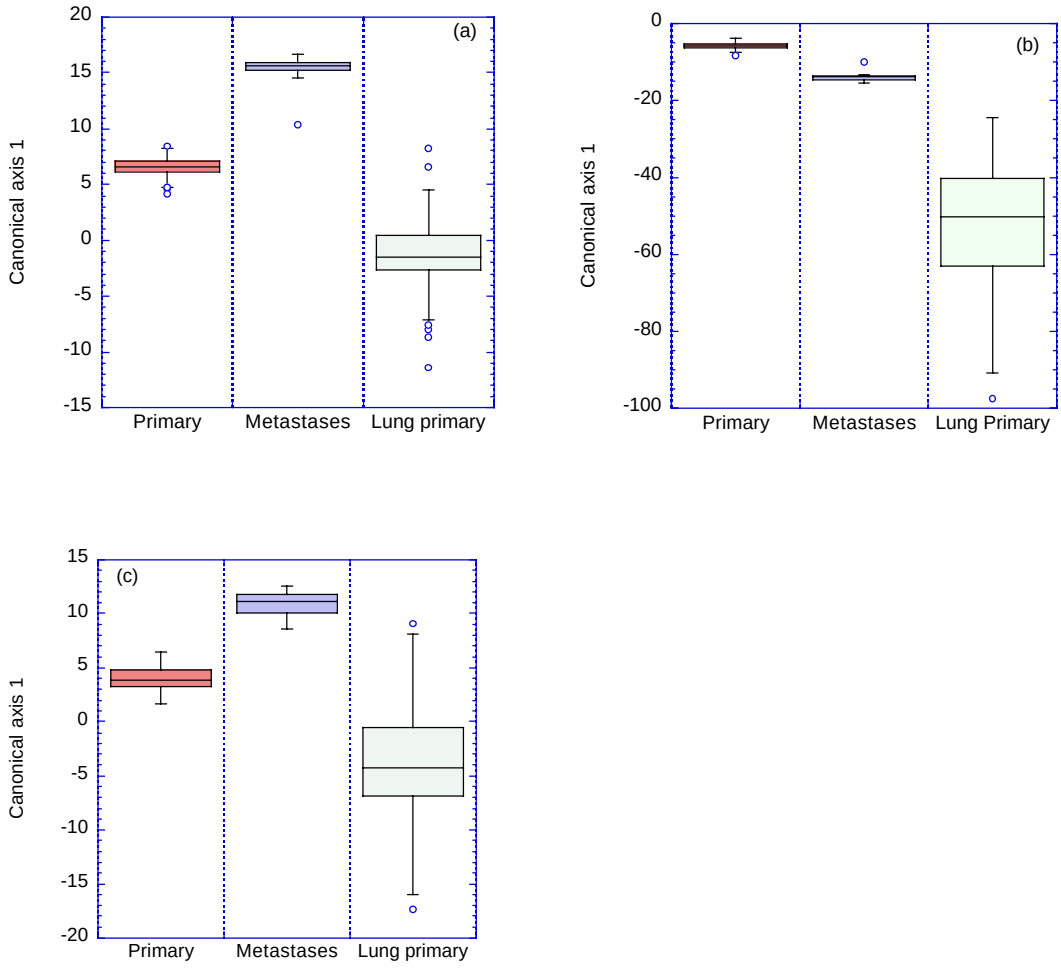


Fig. S.7. Boxplots of original data of primary tumors and metastases from 128-gene data set used to find the canonical transformation to canonical space. We also show application of the transformation to lung primary tumor data from 169-gene data set. (a) 34-gene transformation found from Criterion 6: Maximum correlation. (b) 46 gene transformation found from Criterion 1: Sensitivity of Wilks ratio. (c) 30-gene transformation found from BW-BECA.

Section XVI. Seven-group (six primary groups, one metastases group) canonical analysis of 128-gene Primary/Metastases data of Ramaswamy et al. (2003).

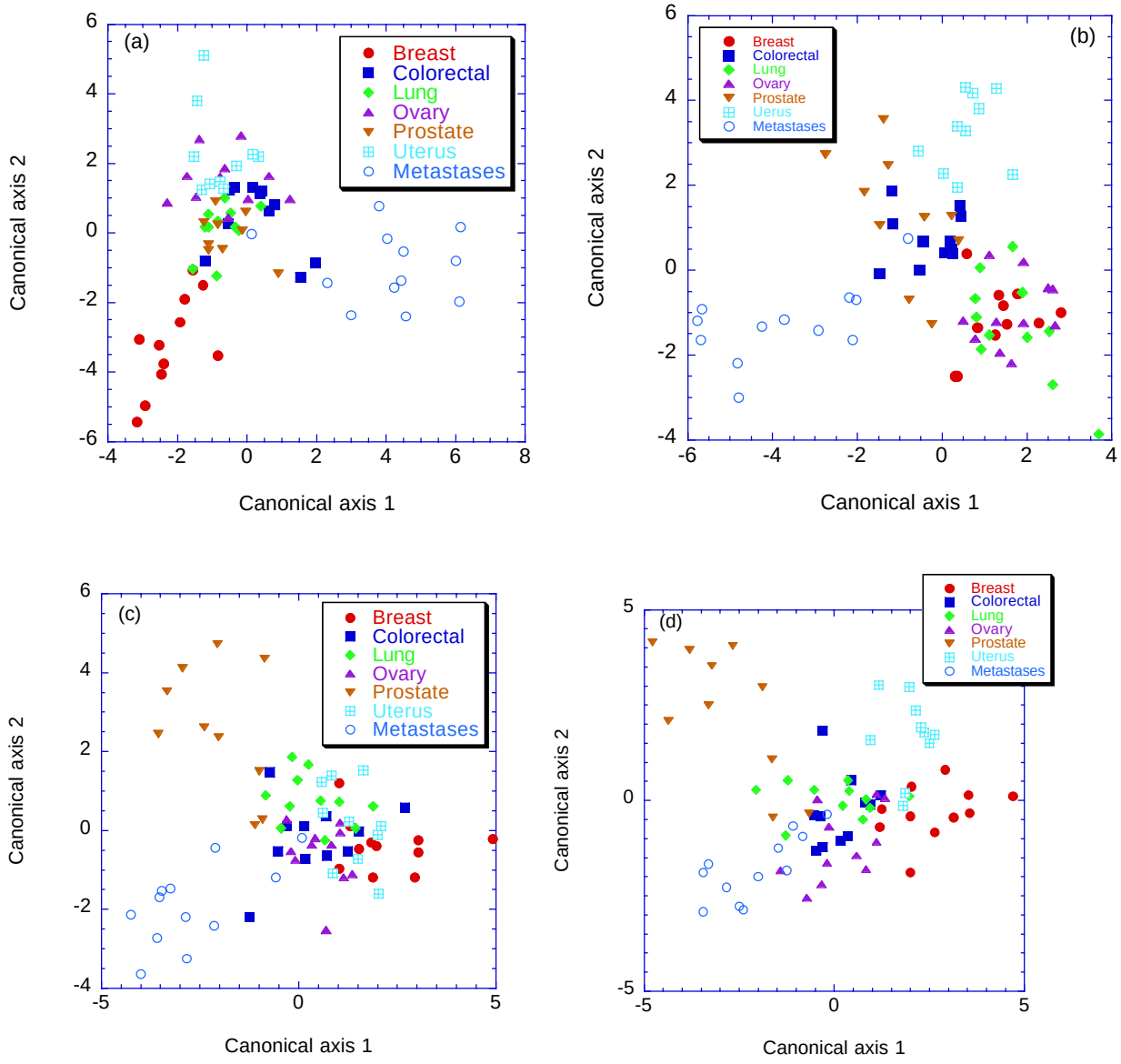


Fig. S.8. See caption next page.

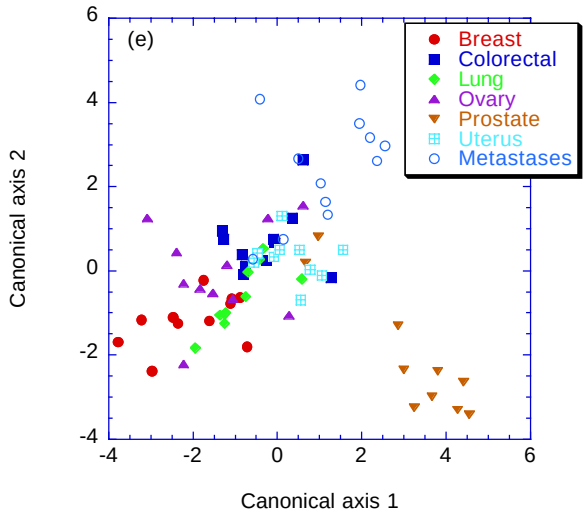


Fig. S.8. Scatter plots on canonical axis 1 and canonical axis 2 of seven group backward elimination CA's for selected results. (a) Criterion 4: (Cumulative probability of coefficients), 16 genes, (b) Criterion 5: (Absolute sensitivity of Wilks ratio), 16 genes, (c) Criterion 6: (Maximum correlation with canonical axes), 14 genes, (d) Criterion 7: (Total correlation with axes), 16 genes, (e) BW-BECA 12 genes.

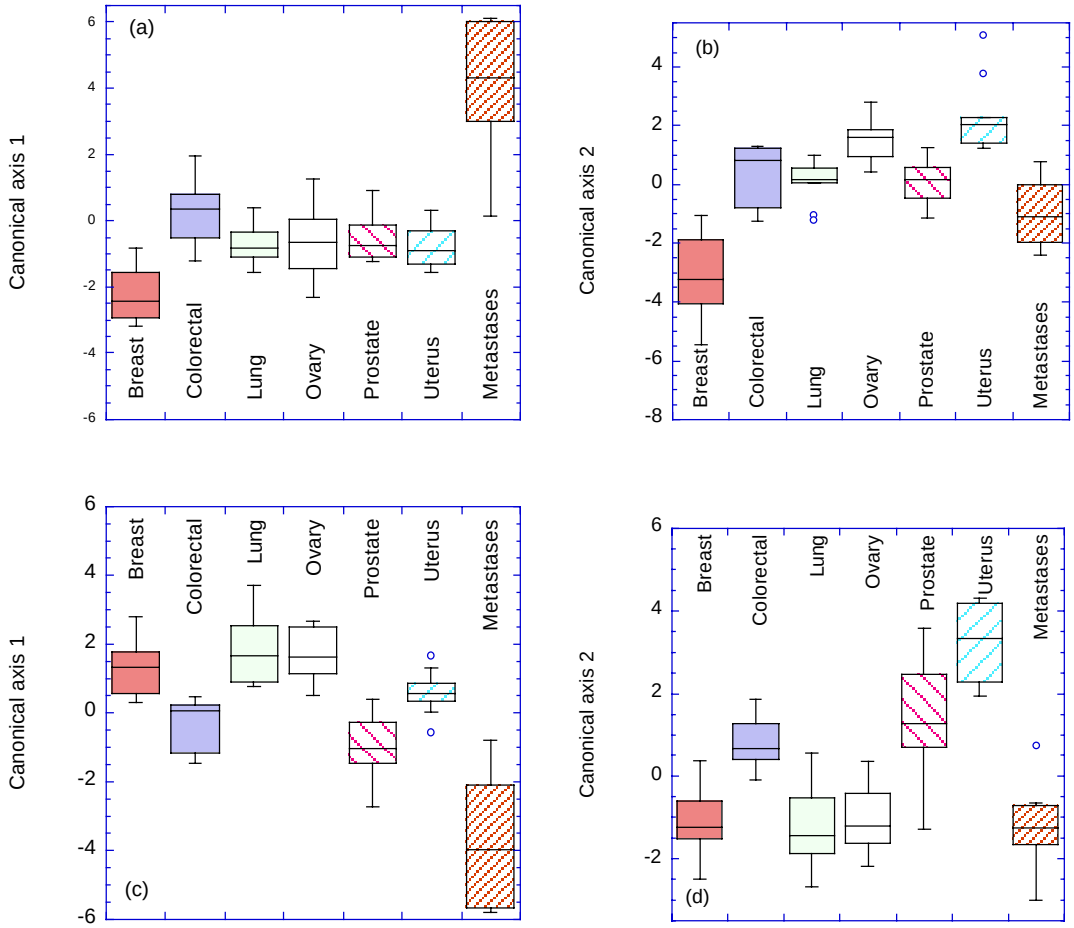


Fig. S.9. See caption next page.

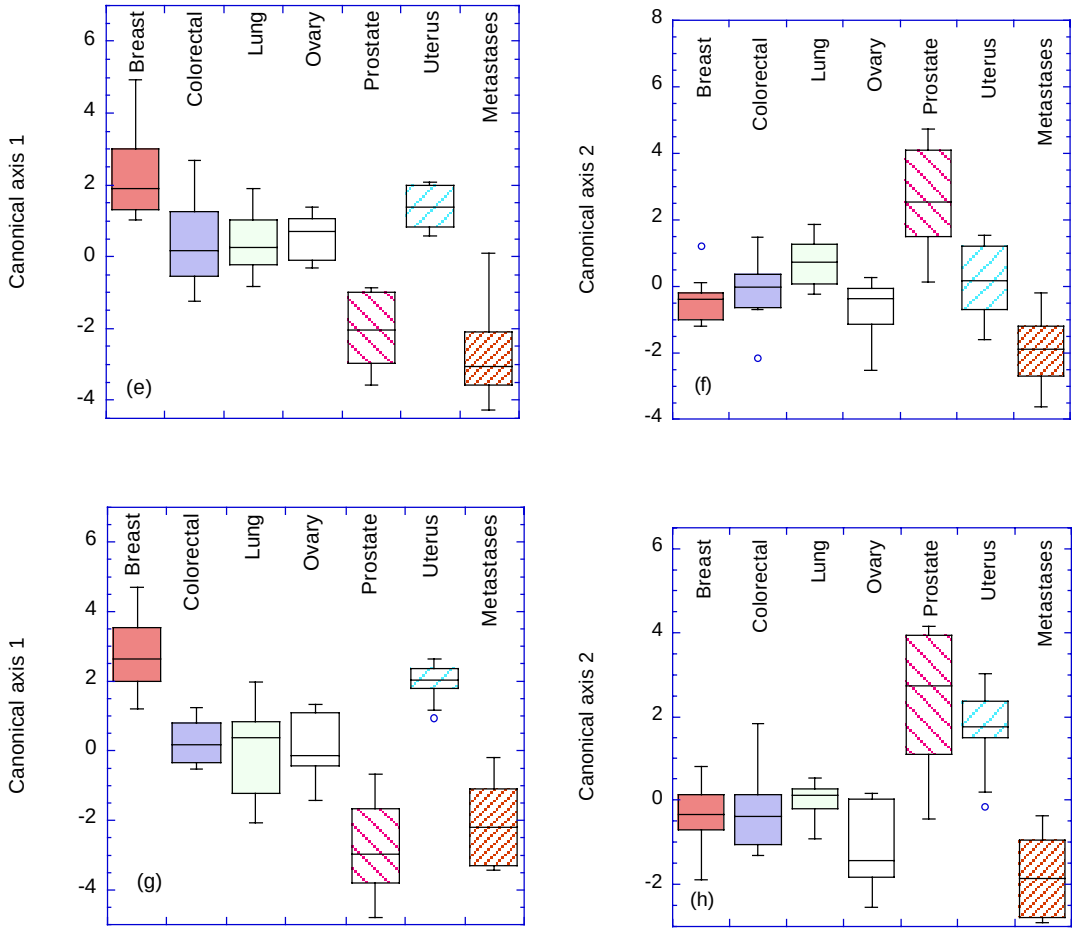


Fig. S.9. Boxplots of seven group canonical analysis from backward elimination procedure. Canonical axis 1 and canonical axis 2 are plotted in Figs. S6a, S6c, S6e, and S6g and Figs. S6b, S6d, S6f, and S6h, respectively. (a) and (b) 16-gene, Criterion 4: Cumulative probability of eliminated variable. (c) and (d) 16-gene, Criterion 5: Absolute sensitivity of Wilks ratio. (e) and (f) 14-gene, Criterion 6: Maximum correlation. (g) and (h) 16 gene, Criterion 7: Total correlation.

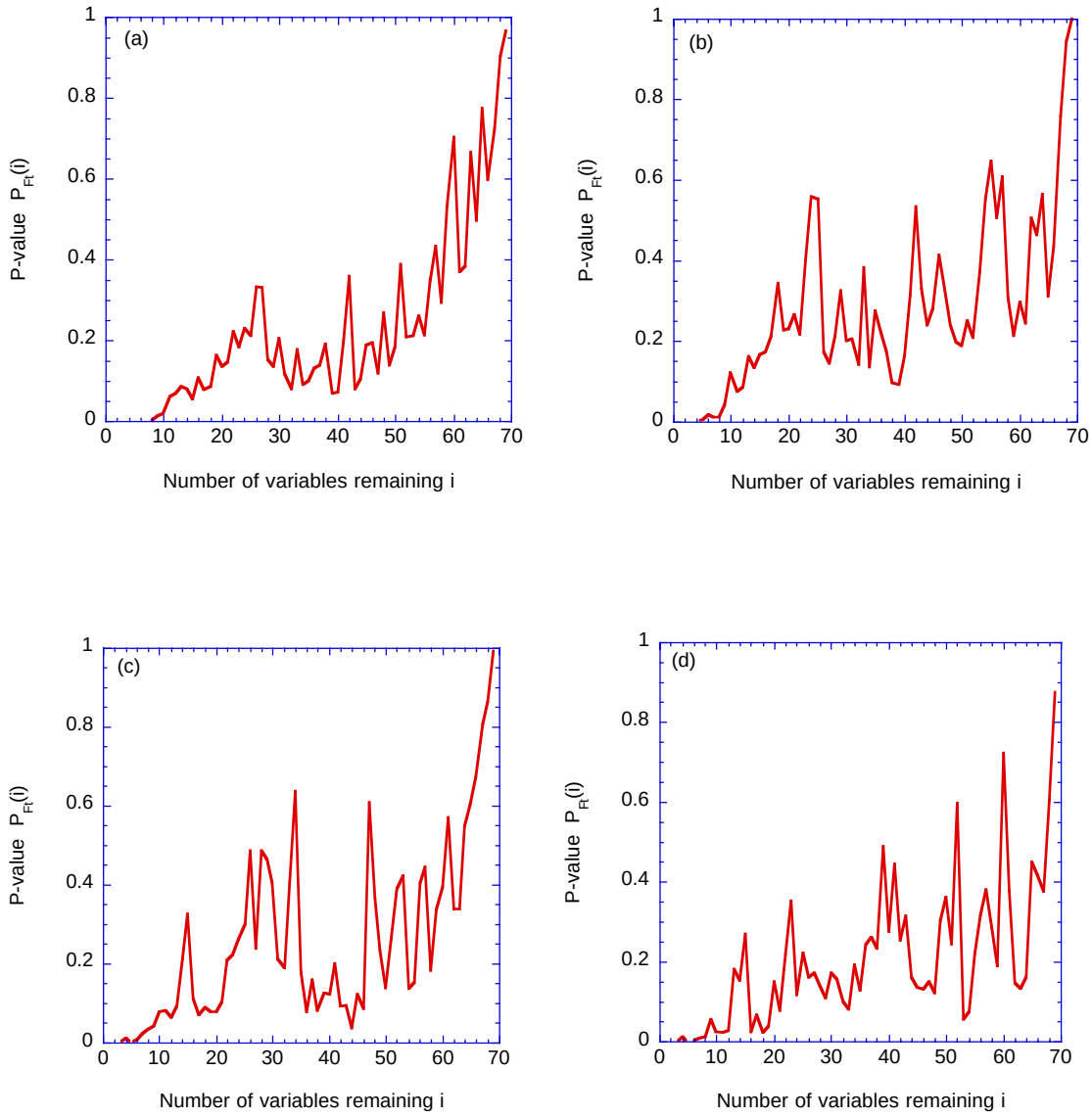


Fig. S.10. See caption next page.

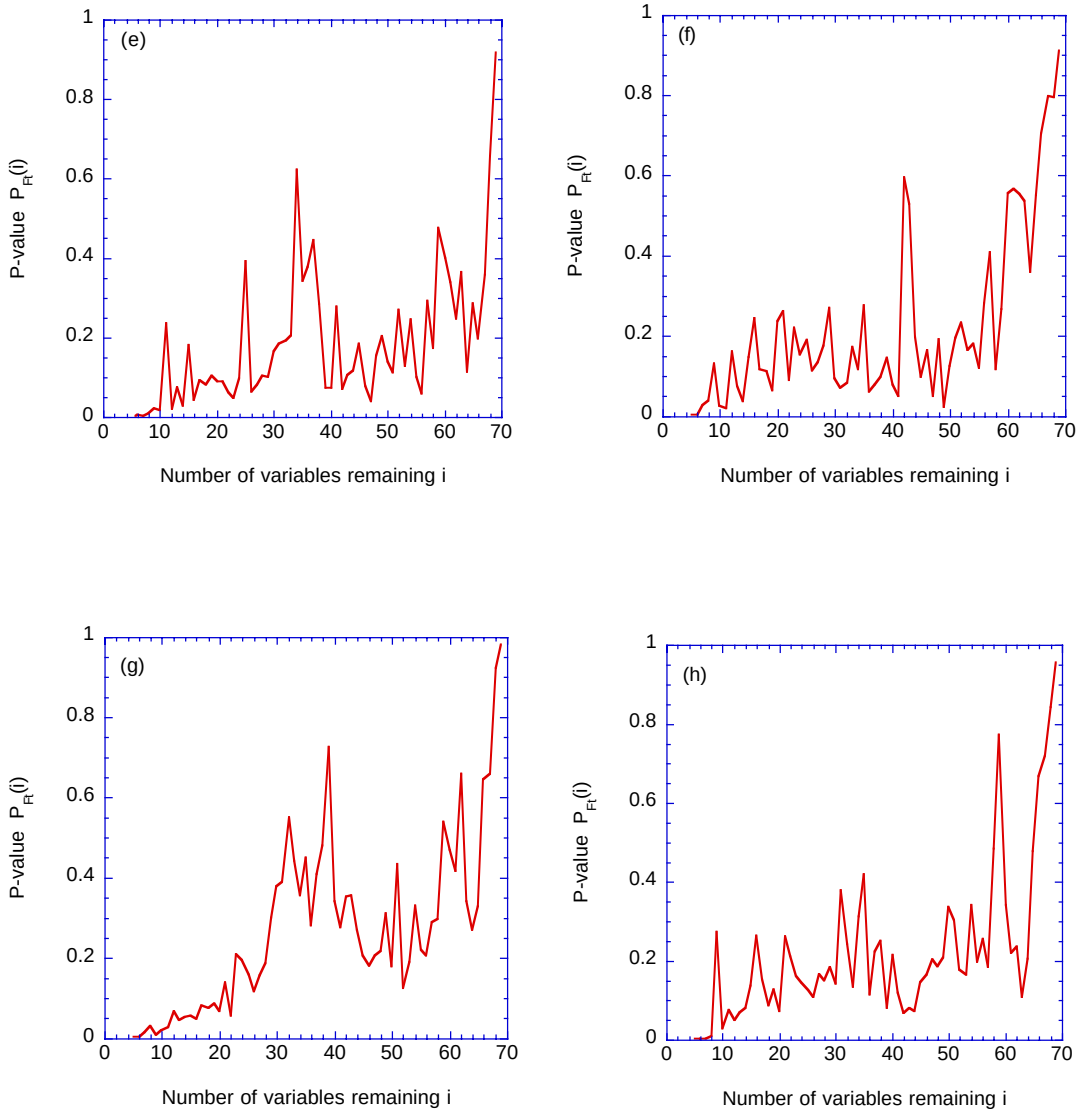


Fig. S.10. *P*-values of Rao-Hawkins-McHenry test ($P_{Ft}(i)$ plotted against *i.* We plot the probability of error if we reject the null hypothesis that the coefficient of eliminated variable is zero. These are graphs are for the 128-gene, seven group case of primary and metastatic tumors. The seven groups are six primary groups (breast, colorectal, lung, ovary, prostate, and uterus) and one metastases group. Data for canonical analyses are from Ramaswamy et al. (2003). (a) criterion 1, (b) criterion 2, (c) criterion 3, (d) criterion 4, (e) criterion 5, (f) criterion 6, (g) criterion 7, and (h) BW-BECA (conventional).

Table S.9. Eigenvalues and significance of sets of associated eigenrvectors for canonical analysis of 16 genes found in backward elimination based on criterion 4: (Cumulative probability of coefficients) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002.

Eigenvalue number	Eigenvalue	Significance of eigenvectors	χ^2	Degrees of freedom	*P*-values of χ^2
1	3.99	$\mathbf{e}^1$ through $\mathbf{e}^6$	274.7	96	3.96E-19
2	2.94	$\mathbf{e}^2$ through $\mathbf{e}^6$	171.8	75	1.46E-09
3	0.787	$\mathbf{e}^3$ through $\mathbf{e}^6$	84.1	56	0.00892
4	0.528	$\mathbf{e}^4$, $\mathbf{e}^5$, $\mathbf{e}^6$	46.9	39	0.179
5	0.230	$\mathbf{e}^5$, $\mathbf{e}^6$	19.8	24	0.709
6	0.107	$\mathbf{e}^6$	6.53	11	0.836

Table S.10. Transformation vectors from gene-variables to canonical axes 1 and 2 for canonical analysis of 16 genes found in backward elimination based on criterion 4: (Cumulative probability of coefficients) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002.

Affymetrix Hu6800/Hu35KsubA ID	$\mathbf{e}^1$	$\mathbf{e}^2$
X82494_at	0.006884	-0.006973
AA093131_at	-0.000241	-0.004587
AA171913_at	-0.002699	-0.004506
U45974_at	0.000112	-0.004557
X94612_at	-0.002056	0.013161
AA486831_s_at	0.000963	0.001844
M64497_at	-0.002408	0.001700
RC_AA195031_at	0.001796	-0.000117
HG3242-HT3419_s_at	0.002040	0.007323
RC_AA411819_at	-0.001357	0.003129
Z14244_at	0.002670	0.001699
RC_AA256996_at	-0.000318	-0.005484
RC_AA449951_at	-0.002753	-0.004342
RC_AA608850_at	0.002247	0.001565
RC_AA100089_at	-0.001771	-0.002709
RC_AA400410_at	0.002671	0.004984

Table S.11. Significance of group separation for canonical analysis of 16 genes found in backward elimination based on criterion 4: (Cumulative probability of coefficients) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002. *P*-values in rejecting the null hypothesis that the groups means are equal is at intersection of the two groups. This is for separation on the first canonical axis.

Group	Primary breast	Primary colorectal	Primary Lung	Primary Ovary	Primary Prostate	Primary Uterus
Primary Colorectal	0.08					
Primary Lung	0.89	1.00				
Primary Ovary	0.84	1.00	1.00			
Primary Prostate	0.82	1.00	1.00	1.00		
Primaty Uterus	0.94	0.99	1.00	1.00	1.00	
Metastases	7.8E-12	5.4E-05	7.7E-08	1.3E-07	4.3E-07	1.1E-07

Table S.12. Significance of group separation for canonical analysis of 16 genes found in backward elimination based on criterion 4: (Cumulative probability of coefficients) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002. *P*-values in rejecting the null hypothesis that the groups means are equal is at intersection of the two groups. This is for separation on the second canonical axis.

Group	Primary breast	Primary colorectal	Primary Lung	Primary Ovary	Primary Prostate	Primary Uterus
Primary Colorectal	2.3E-04					
Primary Lung	1.4E-03	1.00				
Primary Ovary	3.5E-07	1.00	0.94			
Primary Prostate	2.7E-03	1.00	1.00	0.93		
Primaty Uterus	5.8E-09	0.61	0.31	1.00	0.32	
Metastases	0.19	0.89	0.99	0.07	0.99	2.2E-03

Table S.13. Significance of group separation for canonical analysis of 16 genes found in backward elimination based on criterion 4: (Cumulative probability of coefficients) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002. *P*-values in rejecting the null hypothesis that the groups means are equal is at intersection of the two groups. This is for separation on the third canonical axis.

Group	Primary breast	Primary colorectal	Primary Lung	Primary Ovary	Primary Prostate	Primary Uterus
Primary Colorectal	1.00					
Primary Lung	0.89	1.00				
Primary Ovary	1.00	0.94	0.65			
Primary Prostate	0.23	0.98	1.00	0.08		
Primaty Uterus	1.00	1.00	0.90	1.00	0.26	
Metastases	1.00	1.00	0.99	1.00	0.54	1.00

There is no significant group separation of groups on the third axis Thus we ignore any inference from the tests on the axes that the third axes might be significant.

Table S.14. Eigenvalues and significance of sets of associated eigenrvectors for canonical analysis of 16 genes found in backward elimination based on criterion 5: (Absolute sensitivity of Wilks ratio) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002.

Eigenvalue number	Eigenvalue	Significance of eigenvectors	χ^2	Degrees of freedom	P-value of χ^2
1	3.80	$\mathbf{e}^1$ through $\mathbf{e}^6$	289.9	96	2.38E-21
2	2.80	$\mathbf{e}^2$ through $\mathbf{e}^6$	189.5	75	7.01E-12
3	1.32	$\mathbf{e}^3$ through $\mathbf{e}^6$	104.1	56	1.01E-04
4	0.473	$\mathbf{e}^4$, $\mathbf{e}^5$, $\mathbf{e}^6$	50.2	39	0.109
5	0.291	$\mathbf{e}^5$, $\mathbf{e}^6$	25.4	24	0.386
6	0.151	$\mathbf{e}^6$	9.02	11	0.620

Table S.15. Transformation vectors from gene-variables to canonical axes 1 and 2 for canonical analysis of 16 genes found in backward elimination based on criterion 5: (Absolute sensitivity of Wilks ratio) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002.

Affymetrix Hu6800/Hu35KsubA ID	$\mathbf{e}^1$	$\mathbf{e}^2$	$\mathbf{e}^3$
D00654_at	-0.000096	-0.000536	-0.000530
HG110-HT110_s_at	-0.002557	0.001717	0.001197
AFFX-M27830_M_at	-0.000028	0.000108	0.000268
X82494_at	-0.006096	-0.006272	-0.000661
S80437_s_at	-0.000879	-0.000125	0.000137
D31883_at	0.000802	-0.000523	-0.000068
Z74615_at	0.000271	-0.000161	-0.000027
RC_AA447110_at	-0.002448	-0.007149	0.000475
D17408_s_at	-0.000543	0.002284	0.000316
J02783_at	0.000066	0.000453	0.000120
RC_AA600140_at	0.000223	0.000737	-0.000393
RC_AA252812_at	-0.002253	-0.015205	-0.000546
U09851_s_at	0.008552	-0.008028	-0.002839
AA412620_s_at	-0.003713	0.005177	0.004497
AF001548_rna1_at	-0.000253	0.000396	-0.000017
U48959_at	0.000913	-0.000644	0.000411

Table S.16. Significance of group separation for canonical analysis of 16 genes found in backward elimination based on criterion 5: (Absolute sensitivity of Wilks ratio) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002. *P*-values in rejecting the null hypothesis that the groups means are equal is at intersection of the two groups. This is for separation on the first canonical axis.

Group	Primary breast	Primary colorectal	Primary Lung	Primary Ovary	Primary Prostate	Primary Uterus
Primary Colorectal	0.78					
Primary Lung	1.00	0.39				
Primary Ovary	1.00	0.44	1.00			
Primary Prostate	0.22	1.00	0.051	0.06		
Primaty Uterus	1.00	1.00	0.99	0.99	0.89	
Metastases	1.9E-08	4.9E-04	1.5E-09	2.0E-09	0.03	4.2E-06

Table S.17. Significance of group separation for canonical analysis of 16 genes found in backward elimination based on criterion 5: (Absolute sensitivity of Wilks ratio) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002. *P*-values in rejecting the null hypothesis that the groups means are equal is at intersection of the two groups. This is for separation on the second canonical axis.

Group	Primary breast	Primary colorectal	Primary Lung	Primary Ovary	Primary Prostate	Primary Uterus
Primary Colorectal	0.46					
Primary Lung	1.00	0.32				
Primary Ovary	1.00	0.65	1.00			
Primary Prostate	0.11	1.00	0.06	0.21		
Primaty Uterus	3.6E-06	0.10	1.4E-06	1.1E-05	0.53	
Metastases	1.00	0.34	1.00	1.00	0.07	1.3E-06

Table S.18. Significance of group separation for canonical analysis of 16 genes found in backward elimination based on criterion 5: (Absolute sensitivity of Wilks ratio) for seven tumor groups (six primary types, one metastases group) in 128-gene data set of Ramaswaamy et al. 2002. Probability of error in rejecting the null hypothesis that the groups means are equal is at intersection of the two groups. This is for separation on the third canonical axis.

Group	Primary breast	Primary colorectal	Primary Lung	Primary Ovary	Primary Prostate	Primary Uterus
Primary Colorectal	1.00					
Primary Lung	1.00	1.00				
Primary Ovary	1.00	1.00	1.00			
Primary Prostate	0.07	0.01	0.26	0.01		
Primaty Uterus	0.98	1.00	0.84	1.00	0.0003	
Metastases	1.00	1.00	0.99	1.00	0.0022	1.00

Section XVII. Lung data. Differences in groups of recurrent/non-recurrent tumors

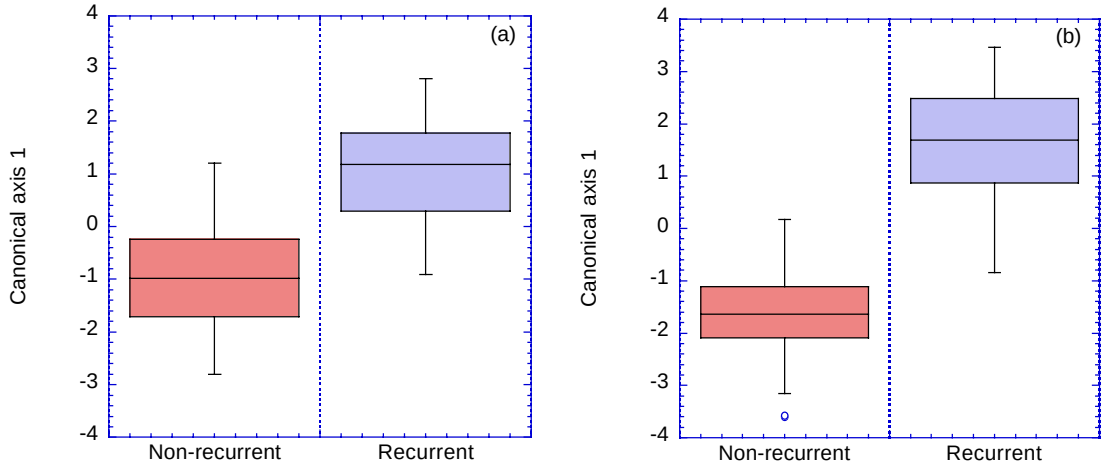


Fig. S.11. Result of backspace elimination canonical analysis at nine genes and eighteen genes in (a) and (b), respectively, for lung primary data from 169-gene data set. The groups were those designated as recurrent or non-recurrent tumors as determined by clinical observation. The data was from Ramaswamy et al. (2003). Both figs were from Criterion 5: Absolute sensitivity of Wilks ratio.

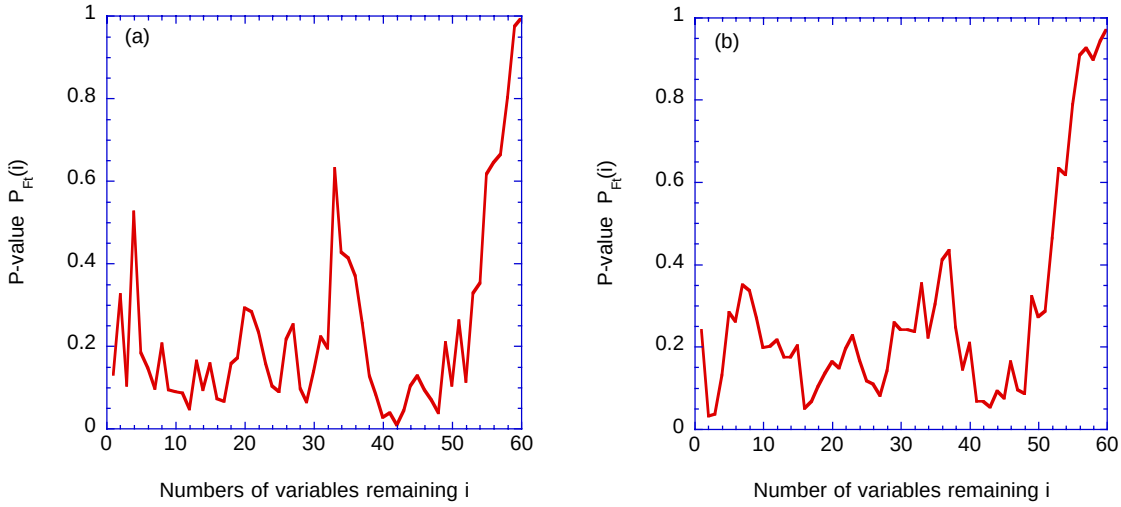


Fig. S.12. See caption next page.

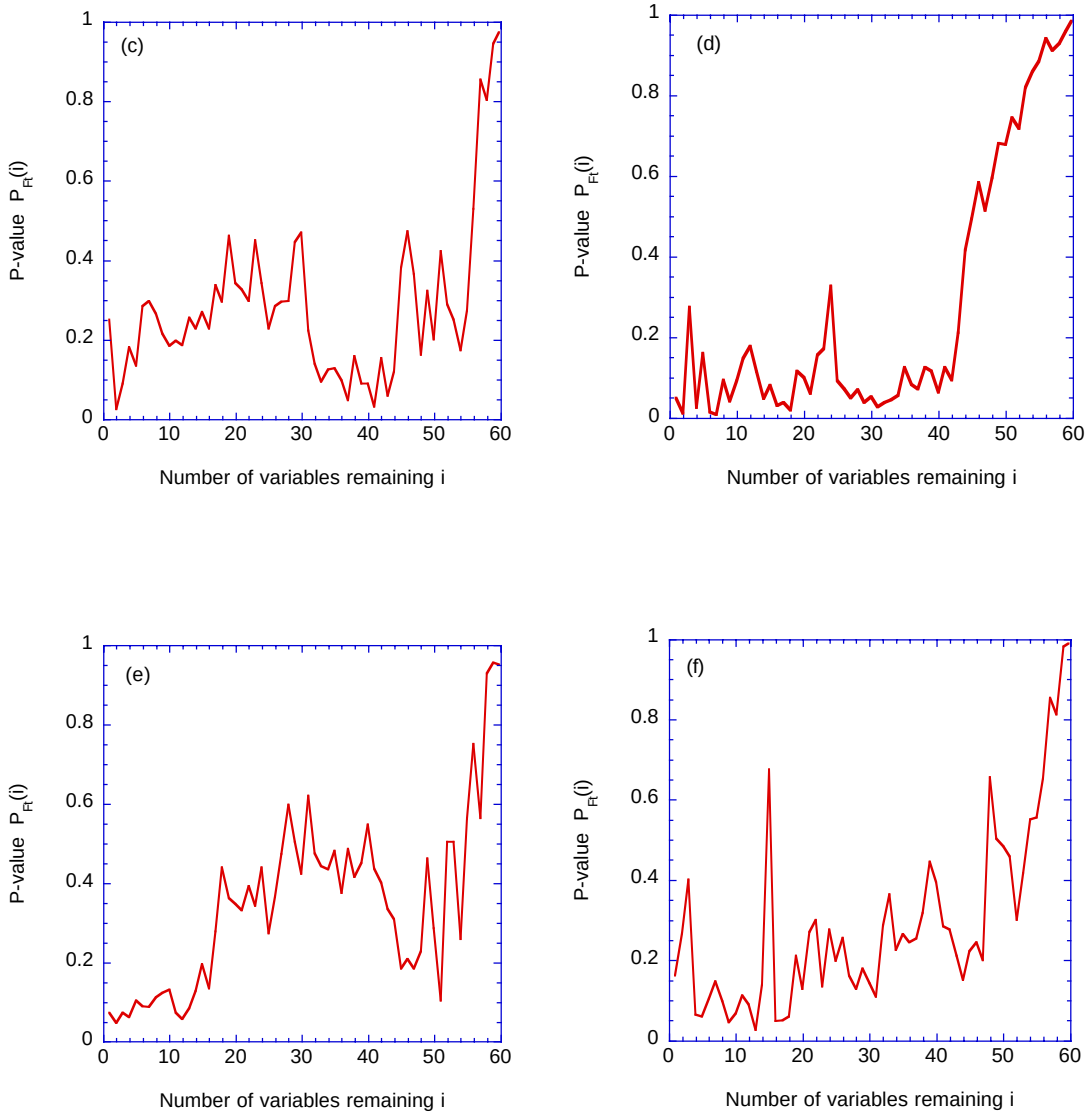


Fig. S.12. Significance of Rao-Hawkins-McHenry test ($P_{Ft}(i)$ plotted against i. We plot the P-value for rejecting the null hypothesis that the coefficient of the eliminated variable is zero. These graphs are for the 169-gene, two group case of lung tumors classified as either recurrent or non-recurrent. These classifications result from clinical observations. Data for canonical analyses are from Ramaswamy et al. (2003). : (a) Criterion 1: Sensitivity of Wilks ratio, (b) Criterion 3: Probability of coefficients (Significance of eliminated variable) and probability of Wilks ratio, (c) Criterion 4: Cumulative probability of coefficients, (d) Criterion 5: Absolute sensitivity of Wilks ratio, (e) Criterion 6-7: Correlation with canonical axis, and (f) BW-BECA.

REFERENCES

Campbell, S.L., C.D. Meyer, Jr. 1979. Generalized inverses of linear transformations. Pitmsn Press.

Dudoit, S., J. Fridlyand, T.P. Speed. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Statistic. Assoc. 97:77-87.

Harris, R.J. 2001. A primer of multivariate statistics. 3rd ed. Lawrence Erlbaum Assoc.: Mahwah, NJ.

Hawkins, D.M. 1976. The subset problem in multivariate analysis of variance. J. Royal Statist. Soc. B 38:132-139.

Kercher, J.R., R.G. Langlois, B.A. Sokhansanj, C.F. Melius, J.N. Quong, F.P. Milanovich, B.W. Colston, Jr., K.W. Turteltaub, A.A. Quong. 2004a. Variable selection in canonical analysis of gene- and protein-expression data: the special case of two groups. (Submitted)

Kercher, J.R., J. N. Quong, A. A. Quong. 2004b. Variable selection in canonical analysis of gene- and protein-expression data: the general case for multiple groups. (Submitted).

Krzanowski, W.J. 2000. Principles of multivariate analysis. Oxford University Press.

Mardia, K.V., J.T. Kent, J.M. Bibby. 1979. Multivariate analysis. Academic Press.

McHenry, C.E. 1978. Computation of a best subset in multivariate analysis. Appl. Statist. 27:291-296.

Morrison, D.F. 1990. Multivariate statistical methods. 3rd ed. McGraw-Hill.

Ramaswamy, S., K.N. Ross, E.S. Lander, T.R. Golub. 2003. A molecular signature of metastasis in primary solid tumors. Nature Geneteics 33:49-54.

Rao, C.R. 1965. Linear statistical inference and its applications. John Wiley, New York.

Rao, C.R. 1970. Inference on discriminant function coefficients. P. 587-602. *In* Essays on Probability and Statistics (R.C. Bose, I.M. Chakravarti, P.C. Mahalanobis, C.R. Rao, K.J.C. Smith, eds) University of North Carolina Press:Chapel Hill, NC.

Schott, J.R. 1997. Matrix analysis for statistics. John Wiley & Sons : New York.

Seal, H. 1964. Multivariate statistical analysis for biologists. Wiley.

Aim 2

Characterize the normal range for candidate host response surrogate markers in humans and animals (year 2 – 3)

Limited Dynamic Range of Immune Response Gene Expression Observed in Healthy Blood Donors Using RT-PCR

Kevin McLoughlin,[1] Ken Turteltaub,[1] Danute Bankaitis-Davis,[2] Richard Gerren,[2] Lisa Siconolfi,[2] Kathleen Storm,[2] John Cheronis,[3] David Trollinger,[2] Dennis Macejak,[2] Victor Tryon,[2] and Michael Bevilacqua[2]

[1]Lawrence Livermore National Laboratory, Livermore, CA, USA; [2]Source Molecular Diagnostics (Source MDx), Boulder, CO, USA; [3]Paradocs Biomedical, Conifer, CO, USA.

The use of quantitative gene expression analysis for the diagnosis, prognosis, and monitoring of disease requires the ability to distinguish pathophysiological changes from natural variations. To characterize these variations in apparently healthy subjects, quantitative real-time PCR was used to measure various immune response genes in whole blood collected from blood bank donors. In a single-time-point study of 131 donors, of 48 target genes, 43 were consistently expressed and 34 followed approximately log-normal distribution. Most transcripts showed a limited dynamic range of expression across subjects. Specifically, 36 genes had standard deviations (SDs) of 0.44 to 0.79 cycle threshold (C_T) units, corresponding to less than a 3-fold variation in expression. Separately, a longitudinal study of 8 healthy individuals demonstrated a total dynamic range (> 2 standard error units) of 2- to 4-fold in most genes. In contrast, a study of whole blood gene expression in 6 volunteers injected with LPS showed 15 genes changing in expression 10- to 90-fold within 2 to 5 h and returning to within normal range within 21 hours. This work demonstrates that (1) the dynamic range of expression of many immune response genes is limited among healthy subjects; (2) expression levels for most genes analyzed are approximately log-normally distributed; and (3) individuals exposed to an infusion of bacterial endotoxin (lipopolysaccharide), show gene expression profiles that can be readily distinguished from those of a healthy population. These results suggest that normal reference ranges can be established for gene expression assays, providing critical standards for the diagnosis and management of disease.

Online address: http://www.molmed.org
doi: 10.2119/2006-00018.McLoughlin

INTRODUCTION

Recent developments in gene and protein expression analysis technology have suggested that gene expression is a key indicator of an individual's pathophysiologic status (1-4). Consequently, clinical application of gene expression technology will vastly improve on the current approaches for monitoring health and disease. Compelling associations between gene expression and disease have been demonstrated in many studies ranging from inflammatory disease to cancer. For instance, studies have pointed to abnormal gene expression in peripheral blood mononuclear cells in lupus patients compared with healthy controls (5,6). Other studies have found differences in gene expression patterns between cancerous liver or pancreatic tissue and nontumor liver and pancreatic tissues (7,8). Additionally, gene expression profiling of breast tumor biopsy tissue correlated with therapeutic response to treatment (9). Results from these studies demonstrate that measurements of gene expression can be used in the diagnosis and monitoring of disease. However, a key requirement for clinical application of gene expression technology is distinguishing between natural variations in gene expression among healthy subjects and changes associated with a disease condition. The establishment of a normal range of expression for a particular population is required as a "reference range" (10).

Immune function is controlled by a network of molecular and cellular pathways. It is well recognized that suppressed immune responses (for example, immunosuppressive therapies and AIDS) or excessive responses (for example, acute respiratory distress syndrome and autoimmunity) can contribute to disease. Thus, homeostatic control and tight regulation of responses are fundamental characteristics of the immune system. For example, in the absence of disease, body temperature remains relatively constant within an individual, suggesting that the body strives to hold its temperature close to a defended set point. During a response to infection, the inflammatory cytokines interleukin-1, interleukin-6, and tumor necrosis factor are released into the blood and bind with receptors in the hypothalamus, resulting in fever (11). However, immune cells also manufacture and release factors, such as interleukin-1 receptor antagonist and interleukin-10, that counteract the effects of pro-inflammatory cytokines and reduce body temperature (12,13). As a result, body temperature rises only moderately during many fever episodes, and returns to its previous set point upon clearance of the infection.

Address correspondence and reprint requests to *Danute M. Bankaitis-Davis, Source MDx, 2500 Central Ave., Suite H-2, Boulder, CO 80301. Phone: (303) 385-2721; fax: (303) 385-2750; e-mail: bunki.davis@sourcemdx.com*
Submitted March 12, 2006; accepted for publication June 7, 2006.

This and other evidence (14) imply that inflammatory/immune genes may be tightly regulated. It is further hypothesized that immune system homeostasis would be reflected in a narrow range of expression levels or set points for key molecules in these pathways among healthy subjects.

In certain gene expression studies, reproducible patterns in subsets of genes have been noted in normal tissues (15-18). The majority of these studies have used microarrays to explore the patterns of expression in isolated blood cell fractions (15,18) or other target tissues, including retina (16) and skin (17). Some studies (16,19) have used replicate arrays to assess the relative contributions of technical and biological factors to the overall variation in measurement values. The results show interindividual variation for gene expression, as well as variation over time within an individual. In addition, gene expression can be sensitive to sources of technical variability, such as time after phlebotomy and method of RNA isolation (20-23). Even within a platform, such as microarray, considerable divergence is reported (24).

In recent years, quantitative real-time (QRT) PCR has emerged as an effective and reproducible tool for transcript analysis (25). It measures relative abundances through PCR-based synthesis of target gene amplicons and activation of target-specific fluorescent probes. The amount of fluorescence generated during the exponential amplification phase provides robust comparative abundance measurements for different amplicons in the same or different wells (25). Whole blood contains representative populations of all the mature cells of the immune system as well as secretory proteins associated with cellular communications (26). The earliest observable changes of cellular immune activity are altered levels of gene expression within the various immune cell types (27). Therefore, QRT-PCR can be an effective technology for reproducibly quantifying gene expression in whole blood.

In studies reported here, we explored the variation among apparently healthy blood bank donors in the expression of a set of genes involved in immune responses. QRT-PCR was used to measure immune-related gene expression in whole blood samples, using procedures designed to sustain a high level of precision (repeatability and reproducibility). We tested the observed distribution of values to determine if it was consistent with sampling from a log-normal distribution, as has been asserted for many genes (28,29), and computed maximum likelihood estimates for the parameters of this distribution. We used statistical models to estimate the contributions of gender, age, and ethnicity to the overall differences in expression among subjects. By performing replicate measurements on longitudinal samples from a group of 8 donors, we computed relative proportions of variance arising from technical, temporal, and intersubject variability. Finally, to obtain limits for the dynamic range of expression achievable with a strong inflammatory stimulus, we performed time-course measurements for several immune response genes in a group of healthy volunteers challenged with an infusion of the bacterial endotoxin lipopolysaccharide (LPS).

MATERIALS AND METHODS

Donor Selection

Single-time-point blood samples from 131 blood donors satisfying American Red Cross blood bank standards (30) were obtained from 3 individual donor centers operated by Bonfils Blood Center, Denver, CO, USA. The samples were drawn on 3 different days over a 3-month period. Subject ages ranged from 22 to 69 years, with a median age of 44 years; age was not recorded for 61 subjects. Women (n = 64) and men (n = 67) were represented in about equal numbers. Ethnicity was reported as white/non-Hispanic for 109 subjects, Hispanic for 19, African-American for 2, and Asian/Pacific Islander for 1. No subjects in this study showed overt signs of disease that would make them ineligible to donate blood under American Red Cross standards. Because we cannot rule out undetected disease in the subjects, however, we refer to them as "apparently healthy" (18).

In addition, longitudinal samples were drawn from 8 volunteers (3 women, 5 men, age range 23 to 50 years) from the Denver area. Samples were collected from these donors approximately once per month for 6 to 8 months, yielding a total of 58 samples.

Samples from the blood donor subjects were collected under Western Institutional Review Board Study No. 20010324. The studies were also reviewed by the Lawrence Livermore National Laboratory Institutional Review Board. Written informed consent was obtained from all volunteers.

In a separate study, 6 healthy volunteers were injected intravenously over 1 min with a single dose (30 units/kg) of Gram-negative bacterial LPS, according to an approved protocol at Guys Hospital, London, UK. Blood samples were drawn and assayed before the LPS injection (0 h) and 2 and 5 h after LPS injection. Additional blood samples from 3 of 6 subjects (adult male volunteers who signed an informed consent form) were drawn and assayed 21 h after LPS injection. Medical history, physical examination, routine laboratory examination, and electrocardiogram were all normal. Subjects did not use any medication or have any significant illness within 8 weeks of the study.

Sample Handling, Purification of RNA, and Preparation of cDNA

Blood was collected from study subjects by standard phlebotomy methods using a 21-gauge butterfly needle and PAXgene Blood RNA Tubes (no. 762115; Qiagen, Valencia, CA, USA) to stabilize messenger RNA (mRNA) against degradation and prevent induction of new mRNA expression (23). Samples were gently mixed by inversion and sat at room temperature for 2 to 24 h to ensure complete nucleic acid stabilization. Samples

were then frozen at –70 °C and batch-shipped on dry ice in compliance with International Air Transport Association (IATA) shipping regulations.

Total RNA from PAXgene Blood RNA samples was extracted within 30 days of collection using the PAXgene Blood RNA Kit (no. 762134; Qiagen). RNA samples were treated with RNase-free DNase I (no. 79254; Qiagen) for digestion of contaminating genomic DNA, using manufacturer-recommended protocols during the purification process. Purified RNA samples were placed at –80 °C for long-term storage.

First-strand cDNA was synthesized with random hexamer primers using TaqMan Reverse Transcription reagents (N808-0234; Applied Biosystems, Foster City, CA, USA). Approximately 250 ng RNA was added to a prepared reverse-transcription reagent mixture consisting of PCR Buffer II, 1×; $MgCl_2$, 5.5 mM; random hexamers, 2.5 μM; dNTP blend, 2 mM; RNase inhibitor, 40 units; and MultiScribe Reverse Transcriptase, 125 units. Samples were incubated at ambient temperature for 10 min with subsequent incubation at 37 °C for 60 min. After the 37 °C incubation, samples were incubated at 90 °C for 10 min and immediately chilled on ice. Newly synthesized cDNA samples were then placed at –80 °C for storage. Prior to QRT-PCR analysis, each cDNA sample was quality control tested for RNA quantity and quality of target genes using quantitative PCR analysis (QPCR; ABI Prism 7700 Sequence Detection System, Applied Biosystems, Foster City, CA, USA) of the 18S rRNA and β-actin.

QRT-PCR Analysis of Target Genes

Primer/probe reagents were custom-designed to achieve 3 performance criteria: (1) single-gene specificity of amplification as tested by gel electrophoresis, (2) dilutional linearity of amplification performance over 2 orders of magnitude, and (3) optimal amplification efficiency of 100 ± 6%, to yield a 2-fold change in transcript per C_T unit (31). Primer/probe sets were designed to span 90 to 120 base pairs, optimized for robust amplification and specificity, minimization of secondary hybridization, and consistent performance. Quality control testing of reagents and manufactured plates ensured that amplification specificity and efficiency remained within established metrics during storage and new synthesis of nucleotides.

Amplification specificity was tested by QRT-PCR with a custom cDNA standard template of induced whole blood and cell lines. Specificity was determined by the size, number, and DNA sequence of the amplified product. The size and number of amplified products was determined by agarose gel electrophoresis. Amplified products were electrophoresed on a 4% agarose gel to visualize the number of DNA bands present. The molecular weight of each band was determined by comparison to known molecular weight markers (no. PR-G1741; Fisher Scientific, Hampton, NH, USA). The presence of a single DNA band of the correct size suggested specific amplification of the intended gene sequence. In certain cases, the amplified product DNA sequence was compared with the published sequence. Primer/probe amplification of genomic DNA was investigated using purified genomic DNA rather than cDNA as the template for QRT-PCR. The formation of primer dimers and spurious amplification was also investigated using DEPC water as a "no template" control for the QRT-PCR assay.

Amplification efficiency of a primer/probe set was determined by a dilutional linearity assay, using 5 serial dilutions of the standard cDNA template and running PCR reactions on each dilution in replicates of 4. Two or more versions of each target gene primer/probe set were designed and tested to select for both amplification efficiency and specificity. Similarly, each new primer/probe reagent lot was monitored to ensure matched amplification specificity and efficiency to previous primer/probe reagent lots.

Target gene transcripts were analyzed by QRT-PCR for each cDNA preparation using 2× TaqMan Universal PCR Master Mix (no. 4305719; Applied Biosystems) and Source MDx's proprietary primer-probe sets. Reactions were run in sets of 4 replicates per gene (24 gene targets in a 96-well plate) on an ABI Prism 7700 Sequence Detection System. Each well also contained the specific primers and probe set to measure 18S rRNA as an internal control. The amount of cDNA template added to each reaction was held to a relatively narrow range, as determined by the cDNA quality control measurement of 18S RNA.

Data Analysis

The difference between the fluorescence C_T for the target gene and the endogenous control (18S rRNA) is presented as a ΔC_T value (C_T of target – C_T of control]. For reference, a ΔC_T of 2 is approximately equivalent to a 4-fold change in the amount of the transcript. For example, at baseline, TGFβ may have a ΔC_T value of 16; after treatment, that ΔC_T value may increase to 18. This change represents a 2 ΔC_T difference or a decrease of 75% (1/4). The C_T reporting system and estimation of relative gene expression are well described in the literature (32).

C_T values above 37 were not used in the analysis, because they correspond to gene expression levels below the linear range of the assay. Values over this threshold were obtained for varying proportions of samples, depending on the gene and the study population examined. For the single-time-point samples, the mean and SD of the underlying ΔC_T distribution were inferred by maximum likelihood estimation (MLE), under the assumption of a normal distribution, for genes having up to 50% of their C_T values over the threshold. Distribution parameters and dynamic ranges were not computed for genes with more than 50% of C_T values greater than 37.

Tests for Normality

Because ΔC_T values are roughly proportional to the logarithm of the corre-

sponding mRNA abundances, we used a combination of analytical methods to test ΔC_T values for each gene for departures from normality.

The Anderson-Darling and Shapiro-Wilk tests were used to test the data against the null hypothesis that the observed values were sampled from a normal distribution, parameterized by the observed mean and standard error. These tests differ in their sensitivity to outliers and in the weight given to central versus outlying values. Smaller *P* values from these tests indicate rejection of the null hypothesis, i.e., deviation from normality.

We also generated plots of the quantiles of each gene's ΔC_T values against the corresponding quantiles of a standard normal distribution (Q-Q normal plots), together with histograms and normal density curves, to graphically characterize their deviations from normality.

Linear Mixed-Effect Model Analysis

Previous reports on longitudinal gene expression data sets (16,19) suggest that, for many genes, expression levels in repeated samples from the same subject are relatively stable compared with interindividual differences, even when the repeat samples are separated by time periods of several weeks. To quantify the relative magnitudes of intersubject versus temporal and technical variability in apparently healthy, untreated subjects, we fitted a linear mixed-effects (LME) model to the longitudinal study data. In this data set, each ΔC_T measurement was associated with a gene *g*, subject *i*, sample index *j*, and replicate *k*. An LME model for these data is described by equation 1:

$$(\Delta C_T)_{gijk} = \alpha_g + u_{gi} + \beta_{gj} + v_{gij} + \varepsilon_{gijk} \quad (1)$$

where α_g is an intercept term dependent on the gene only, u_{gi} is a random effect due to intersubject variability, β_{gj} is a fixed effect due to systematic variations in processing affecting all samples drawn at the same time point, v_{gij} is a random effect representing variability among samples from the same subject, and ε_{gijk} is an error term encompassing all residual sources of variability between replicates. The random effects u_{gi}, v_{gij}, and ε_{gijk} are assumed to be normally distributed with mean zero and variances σ^2_S, σ^2_T, and σ^2_R, respectively. A restricted maximum likelihood (REML) algorithm (33) was used to fit the model parameters α_g, β_{gj}, σ^2_S, σ^2_T, and σ^2_R to the data.

In addition, it is useful to quantify the contributions to intersubject variability arising from subject characteristics such as sex, age, and ethnicity. All 3 of these parameters were recorded for 68 subjects in the single-time-point study. Expression data for these subjects was fitted to the LME model described by equation 2:

$$(\Delta C_T)_{gik} = \alpha_g + \beta_g(G_i, E_i) + \zeta_g(G_i, E_i)A_i + u_{gi} + \varepsilon_{gik} \quad (2)$$

where α_g is an intercept term dependent on the gene only, G_i, A_i, and E_i are the sex, age, and ethnicity of subject *i*, $\beta_g(G,E)$ is a gene-specific offset for the given sex and ethnicity, $\zeta_g(G,E)$ is the slope of a linear age effect depending on both sex and ethnicity, u_{gi} is a random effect due to intersubject variability not explained by age, sex, or ethnicity, and ε_{gik} is an error term encompassing all residual sources of variability between replicate PCR reactions for a given sample. After fitting this model, the percentage contribution of sex, age, and ethnicity effects to the intersubject variance for gene *g* was estimated by equation 3:

$$(PC)_g = 100/(1 + \sigma^2_S / \textstyle\sum_{ik} (((\text{predicted } \Delta C_T)_{gik} - (\text{mean } \Delta C_T)_g)^2)/(N-1)) \quad (3)$$

where *N* is the total number of measurements for gene *g*, σ^2_S is the variance parameter estimated for the distribution of the random subject effects, predicted ΔC_T is the value predicted from the fixed effects portion of equation 2, and mean ΔC_T is computed over all measurements for gene *g*.

All data analyses were performed using the R open source programming environment for statistical computation (34). LME models were programmed using the R package "nlme" (33).

RESULTS

Most Genes Exhibit Limited Dynamic Range of Expression Across Subjects in Single-Time-Point Measurements

A series of studies were undertaken to examine the expression of immune-related gene transcripts in whole blood of apparently healthy subjects. In the largest single-time-point study, blood was collected from 131 blood donors following the American Red Cross donor standards and analyzed for the expression of 48 inflammation- and immune-related gene transcripts. These transcripts encode cell surface molecules, such as CD4, CD14, CD19, and ICAM-1; signaling molecules, such as PTGS2 (COX2), PLA2G7, and NF-κB; cytokines, such as IL-1B and TGFβ; proteinases, such as ELA2; and proteinase inhibitors (see Table 1). The overall range of C_T values for the 48 genes studied is plotted in Figure 1. The bars in the plot encompass the central 90% of the observed values (i.e., they extend from the 5th to the 95th percentiles), whereas the whiskers on either end of the bar extend to the extreme values. For genes with expression levels sampled from a log-normal distribution, the ends of the bars would correspond to 1.64 SD on either side of the mean C_T.

Of the 48 genes profiled in this study, 2 important signals of inflammation, IL6 and CXCL2, lacked detectable expression in most of the apparently healthy subjects, and their C_T values were at or greater than 37. Dynamic ranges and variance components were not computed for these genes. For the remaining 46 genes, the estimated SD of the ΔC_T values ranged from 0.44 to 1.46 and were below 0.792 for 36 of the 46 genes, as shown in Table 1. Thus, the dynamic range of expression extending 2 SD in either direction from the geometric mean was less than $2^2 * °.792$ or a 3-fold change (32). For normally distributed ΔC_T values, this range covers 95.4% of the sample measurements. The distribution of dynamic ranges corresponding to a ±2 SD span is shown in Figure 2. The highest dynamic range observed was 7.53-fold change

Table 1. Genes with detectable expression in healthy blood donor samples, together with statistical summaries of ΔC_T distribution, expression fold changes corresponding to 2 standard deviations of ΔC_T distribution, and *P* values for normality tests.

HUGO Designation	Gene Name and Aliases	*n*	Mean	Median	SD	Fold Change ± 2 SD	Shapiro–Wilk	Anderson–Darling
ADAM17	A Disintegrin and Metalloproteinase Domain 17	129	18.56	18.55	0.63	2.39	0.7512	0.5395
APAF1	Apoptotic Protease Activating Factor 1	131	16.46	16.48	0.54	2.13	2.1E-05	0.0150
C1QA	Complement Component 1, Q Subcomponent, Alpha Polypeptide	128	20.25	20.21	0.92	3.57	0.0939	0.0879
CD14	CD14 Antigen	129	13.92	14.01	0.63	2.41	1.1E-05	8.0E-07
CD19	CD19 Antigen	131	18.19	18.09	0.78	2.94	1.4E-05	1.1E-07
CD4	CD4 Antigen	131	14.80	14.84	0.49	1.98	0.0064	3.8E-04
CD86	CD86 Anitigen; B7–2 Protein	128	17.64	17.68	0.51	2.04	3.1E-05	6.6E-04
CD8A	CD8 Antigen, Alpha Polypeptide, p32	130	15.74	15.72	0.67	2.54	0.0653	0.8402
CXCL1	Chemokine (C–X–C Motif) Ligand 1 (GRO–1)	131	20.01	20.00	0.67	2.53	0.1150	0.1522
CYBB	Cytochrome B–245 Beta Polypeptide	130	13.98	14.02	0.57	2.21	0.0058	0.0542
DPP4	Dipeptidylpeptidase IV (CD26)	131	18.33	18.35	0.61	2.34	0.1253	0.0602
EGR1	Early Growth Response 1	130	20.42	20.49	0.65	2.47	0.0074	0.0013
ELA2	Elastase 2, Neutrophil	126	19.90	19.78	1.29	5.95	2.1E-04	1.4E-04
GCLC	Glutamate–Cysteine Ligase, Catalytic Subunit	128	18.86	18.90	0.64	2.41	5.6E-07	2.9E-05
HMGB1	High–Mobility Group Box 1	130	16.28	16.25	0.69	2.59	0.0055	0.0524
HMOX1	Heme Oxygenase (Decycling) 1	131	16.45	16.50	0.67	2.53	0.0028	0.0045
HSPA1A	Heat Shock Protein 1A, 70 kDa	129	13.83	13.88	0.80	3.01	3.7E-08	1.2E-06
ICAM1	Intercellular Adhesion Molecule 1	131	17.68	17.71	0.55	2.15	0.0969	0.0514
IFI16	Interferon γ-Inducible Protein 16	130	16.75	16.72	0.84	3.20	0.0441	0.1004
IL10	Interleukin 10	75	22.87	22.94	0.75	2.81	9.0E-04	0.0070
IL15	Interleukin 15	129	21.45	21.45	0.70	2.65	0.0051	0.0275
IL18	Interleukin 18 (Interferon γ-Inducing Factor)	130	20.05	20.05	0.54	2.11	0.0517	0.0625
IL18BP	IL–18 Binding Protein	131	16.74	16.72	0.44	1.84	0.2787	0.6132
IL1B	Interleukin 1β	130	16.67	16.67	0.79	2.99	0.0011	0.0200
IL1R1	Interleukin 1 Receptor, Type I	125	21.08	21.06	0.98	3.90	0.5969	0.9508
IL1RN	Interleukin 1 Receptor Antagonist	129	16.88	16.91	0.67	2.54	0.1494	0.1755
IL8	Interleukin 8	97	21.01	20.86	1.46	7.53	0.0321	0.1185
LTA	Lymphotoxin α	114	20.05	19.99	0.65	2.45	3.4E-04	0.0014
MMP9	Matrix Metalloproteinase 9	129	15.91	16.01	1.16	4.97	3.8E-05	1.9E-06
MNDA	Myeloid Cell Nuclear Differentiation Antigen	130	12.54	12.51	0.64	2.44	0.1193	0.1563
MPO	Myeloperoxidase	131	21.20	21.22	0.77	2.92	0.7944	0.7479
MYC	V–myc Avian Myelocytomatosis Viral Oncogene Homolog	130	17.23	17.22	0.63	2.38	0.0685	0.0705
NFKB1	Nuclear Factor of κ Light Polypeptide Gene Enhancer in B Cells 1 (p105)	131	17.38	17.41	0.57	2.20	0.0178	0.0076
PLA2G7	Phospholipase A2, Group VII	126	19.36	19.36	0.69	2.60	0.1485	0.5492
PLAUR	Plasminogen Activator, Urokinase Receptor	131	15.12	15.15	0.58	2.25	0.0275	0.0190
PTGS2	Prostaglandin–Endoperoxide Synthase 2 (COX–2)	126	16.72	16.75	0.61	2.33	0.0505	0.0309
PTPRC	Protein Tyrosine Phosphatase Receptor, 127 Type C (CD45)	11.91	11.96	0.52	2.07	0.0410	0.0165	
SERPINA1	Serine (or Cysteine) Proteinase Inhibitor, Clade A, Member 1 (Alpha 1 Anti–Trypsin)	131	13.26	13.27	0.66	2.50	0.0100	0.0092
SERPINE1	Serine (or Cysteine) Proteinase Inhibitor, Clade E (Ovalbumin), Member 1 (Plasminogen Activator Inhibitor Type 1)	101	22.38	22.44	0.89	3.43	0.0014	0.0015
SERPING1	Serine (or Cysteine) Proteinase Inhibitor, Clade G (C1 Inhibitor), Member 1 (Angioedema, Hereditary)	130	19.20	19.29	1.20	5.25	3.2E-04	6.0E-05
TGFB1	Transforming Growth Factor β 1	130	13.14	13.16	0.44	1.83	0.0115	0.0141
TIMP1	Tissue Inhibitor of Matrix Metalloproteinase 1	131	15.02	15.09	0.57	2.19	1.1E-05	4.2E-06
TLR2	Toll–Like Receptor 2	130	16.07	16.13	0.63	2.38	0.0058	0.0015
TNF	Tumor Necrosis Factor	124	20.67	20.55	0.98	3.90	1.1E-05	3.4E-04
TNFSF5	Tumor Necrosis Factor (Ligand) Superfamily, Member 5 (CD40 Ligand)	131	17.69	17.67	0.63	2.38	1.5E-14	1.5E-10
TNFSF6	Tumor Necrosis Factor (Ligand) Superfamily, Member 6 (Fas Ligand)	126	20.41	20.35	0.74	2.80	4.7E-04	0.0021

n is the number of samples having detectable expression for the gene in at least 3 of 4 replicate RT–PCR reactions. Mean and SD are estimated by maximum likelihood for genes where any replicates fall below the detection threshold ($C_T > 37$).

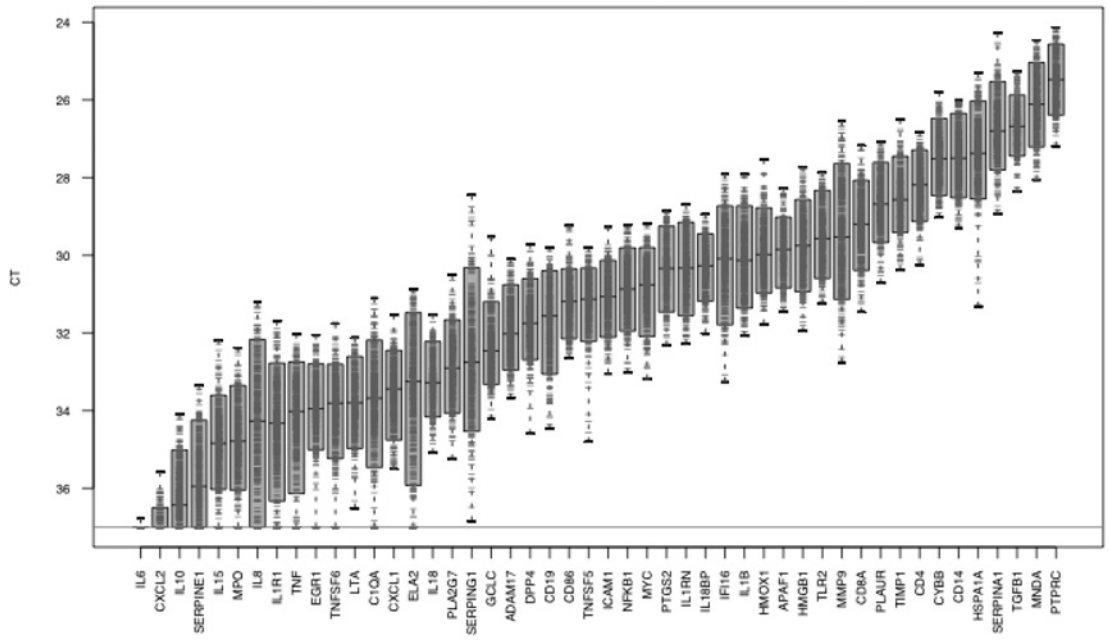


Figure 1. Gene distribution across 131 healthy donors. Range of C_T values for each gene targeted by the panel of 48 primer sets, across 131 single-time samples. Bars span the range from the 5th to the 95th percentile of C_T values for each gene.

units for IL-8. The SDs of ΔC_T values were independent of the mean ΔC_T, indicating that the dynamic ranges did not depend on a gene's expression level.

The Majority of Genes Have Expression Values following Log-Normal Distributions

Commonly used parametric tests for differential gene expression between groups of samples, such as *t* tests and analysis of variance, are based partly on the assumption that the values being compared are sampled from normal distributions. Although it is commonly asserted that transcription levels of many genes are log-normally distributed (28,29), it is important to test this assumption to use such tests for disease diagnosis and detection. The majority of expressed transcripts followed approximately log-normal distributions, according to the Anderson-Darling and Shapiro-Wilk tests (Table 1, Figure 3). The gene most closely following a normal distribution of ΔC_T values was IL1R1 (Figure 3A), with an Anderson-Darling *P* value of 0.945. Among the 46 genes tested, 34 had *P* values greater than 0.001. All genes had unimodal distributions; the deviations from normality involved moderate degrees of left or right skewness, and/or heavy or light tails. Although these departures were not dramatic, they will need to be incorporated into the predicted error rates for diagnostic tests based on expression of these genes.

Of the 48 genes shown in Table 1, the gene deviating most from a normal distribution of ΔC_T values was TNFSF5 (CD40 ligand, Figure 3B), with an Anderson-Darling *P* value of 1.52×10^{-10}. The observed distribution is characterized by a heavy tail and large ΔC_T, suggesting the presence of a subpopulation with an unusually low expression level of this gene.

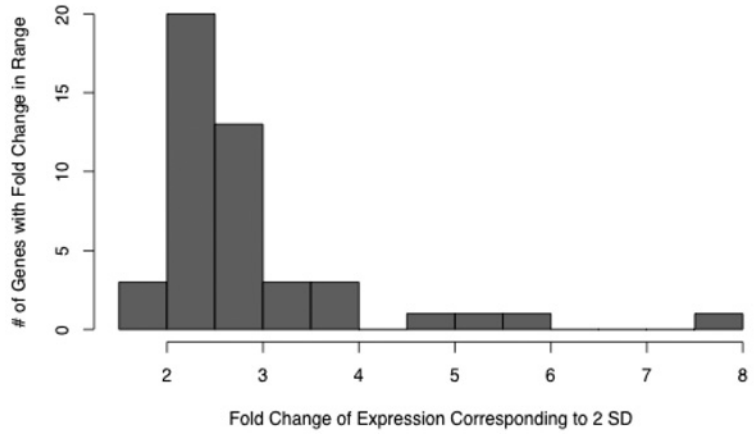


Figure 2. Histogram of dynamic ranges of expression values, expressed as fold changes spanning 2 standard deviations of each gene's ΔC_T values (that is, $2^{-2SD(\Delta C_T)}$).

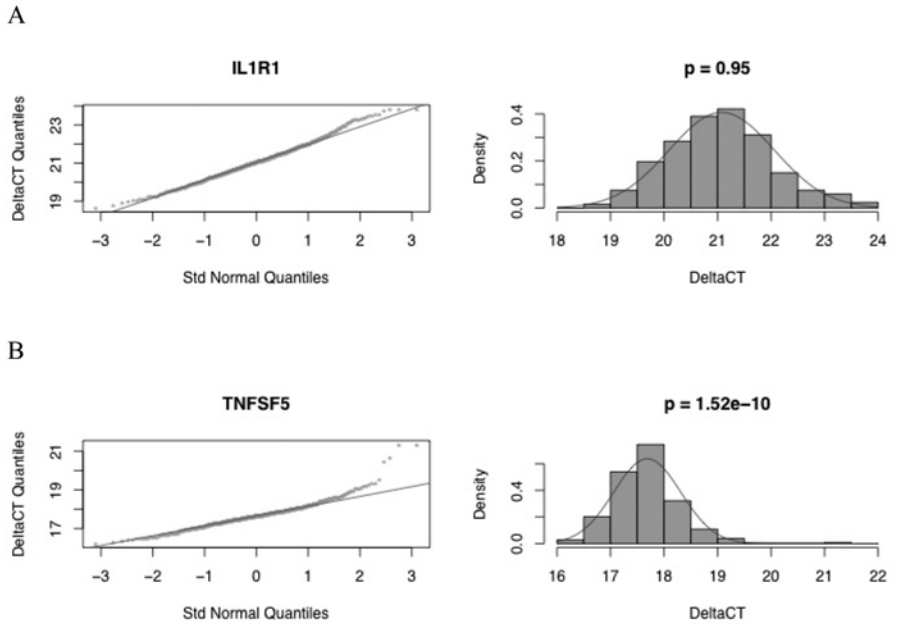


Figure 3. Q-Q normal plots and histograms of ΔC_T values for the genes deviating least and most from a normal distribution (IL1R1 in Figure 3A and TNFSF5 in Figure 3B, respectively), according to the Anderson-Darling test. Unit diagonals and normal density curves are drawn on the Q-Q normal plots and histograms, respectively, for comparison with a normal distribution with the same mean and variance as observed. *P* values computed by the Anderson-Darling normality test are shown above each histogram.

Minor Variations in Expression May Be Based on Sex, Ethnicity, and Age

Table 2 shows the contributions of sex, age, and ethnicity on interindividual variation estimated by the LME model (equation 2). For the 43 genes examined, the observed effects of sex, ethnicity, and age were small. Only 10 genes had contributions from these effects, explaining more than 20% of the intersubject variance; the maximum contribution was only 27.9% for NFKB1. For most genes, sex effects accounted for most of this contribution. Fifteen genes showed significant sex differences (unadjusted *P* value < 0.05), but the largest fold change from women to men was only 1.62 for TNFSF6. Likewise, only moderate ethnicity effects were observed. Five genes (MPO, MYC, TNFSF6, ELA2, and HMGB1) showed significant differential expression between white (non-Hispanic) and Hispanic subjects, with the largest change being a 2.5-fold overexpression of ELA2 in Hispanic women relative to white women.

Age effects were difficult to measure in this data set, due to the markedly different age distributions between the female and male blood donors. Male blood donors had a median age of 53 years, compared with 43 years for females. Therefore, sex and age effects are potentially confounded. The LME model defined in equation 2 addresses the confounding factors by fitting the ΔC_T versus age data to different slopes for each sex/ethnicity combination. According to the LME model, 3 genes (IL18, ELA2, and C1QA) had significant age effects for at least 1 sex/ethnicity combination. For all 3 of these genes, the fitted slopes were markedly different between sexes. For example, age had virtually no effect on IL18 expression in white men, whereas in white women the slope corresponded to a 2-fold increase from age 23 to age 69. Similarly, the fitted slopes suggest dramatic differences in age effects among ethnicities. Overall, the size of the sample is too small to reliably estimate ethnic differences.

Variation of Expression within Subjects Over Time Is Limited

To compare the contributions of intersubject, temporal, and technical components to the overall variation in gene expression, we fitted the LME model (equation 1) to the longitudinal set of measurements described in "Materials and Methods." For this data set, we fitted the model for each of 29 genes with detectable expression in at least 90% of the samples to obtain, for each gene, a set of variance parameters σ^2_S, σ^2_T, and σ^2_R. These are approximate estimates of the contributions to the total variance from intersubject variation, variation among samples taken at different times from each subject, and residual variation between replicate reactions, respectively.

The results of the initial LME model analysis are summarized in Figure 4, which shows the fitted standard error parameters σ^2_S, σ^2_T, and σ^2_R for each gene. For 6 of the 29 genes examined (CD19, TNFSF13B, HMOX1, C1QA, CD8A, and CD4), intersubject variation comprised more than 50% of the total variance of ΔC_T values. For the remaining 23 genes, variation between samples taken at different times was the largest component. However, the magnitude of the temporal variation was limited; the parameter σ_T ranged from 0.36 ΔC_T units for the gene PTPRC to 0.72 ΔC_T units for MMP9. The dynamic ranges corresponding to $2\sigma_T$ ranged from 1.66- to 2.72-fold change units. Because measurements from samples taken over a period of 8 months may be subject to several sources of technical variation (for example, instrument calibration, reagent lots, and variations in sample handling), these ranges can be considered upper limits on the true temporal variation of expression for the genes analyzed.

LPS Stimulation Induces Transient Gene Expression Changes in Excess of Natural Variation

To demonstrate that changes marked beyond the normal reference range occur, gene expression was measured in blood

Table 2. Sex, age, and ethnicity (fixed effect) contributions to intersubject variation for 43 genes, in decreasing order of percentage of variance explained (equation 3).

Gene	Percent of Variance Explained by Sex, Age, and Ethnicity	*P* Values for Effect							Fold Change Corresponding to Effect			Fold Change Corresponding to Age Difference (69 vs. 23 years)			
		Sex	Ethnicity	Sex + Ethnicity	Age	Sex + Age	Ethnicity + Age	Sex + Ethnicity + Age	WM vs. WF	HF vs. WF	HM vs. WF	WF	WM	HF	HM
NFKB1	27.90	0.0022	0.7083	0.1098	0.3335	0.0553	0.1896	0.2572	1.35	-1.04	-1.03	-1.29	-2.48	1.32	1.43
MPO	27.74	0.0128	0.0005	0.0116	0.2260	0.3535	0.8682	0.0909	1.42	1.82	1.31	1.57	1.00	1.43	4.48
IL18	27.46	0.0220	0.6228	0.9333	0.0119	0.0468	0.9150	0.1023	1.25	1.06	1.30	1.97	-1.02	1.88	2.77
MYC	26.91	0.0180	0.0132	0.7298	0.5240	0.1344	0.9658	0.1843	1.31	-1.42	-1.17	-1.21	-2.19	-1.24	1.23
TGFB1	25.88	0.0121	0.0726	0.5024	0.5696	0.0809	0.3451	0.1967	1.22	-1.18	-1.07	-1.13	-1.84	1.21	1.47
TNFSF6	23.90	0.0008	0.0344	0.0123	0.1039	0.0892	0.8499	0.4898	1.62	1.39	1.17	1.78	-1.28	1.61	1.32
LTA	23.87	0.0179	0.8223	0.0131	0.4992	0.1109	0.9253	0.4470	1.32	-1.03	-1.33	-1.22	-2.32	-1.17	-1.27
ELA2	23.65	0.3536	0.0013	0.0998	0.0262	0.2949	0.1695	0.0889	1.24	2.53	1.50	4.21	1.78	1.09	6.94
CD86	21.03	0.0289	0.2781	0.4782	0.1418	0.1509	0.3672	0.3403	1.24	1.13	1.23	1.47	-1.11	2.12	2.42
CD14	20.75	0.0066	0.9157	0.1150	0.1484	0.4972	0.0631	0.6869	1.42	-1.02	-1.05	-1.65	-2.24	1.64	1.71
C1QA	19.11	0.9650	0.2648	0.7856	0.0458	0.1489	0.0286	0.0015	1.01	1.25	1.15	2.51	1.07	-1.89	9.27
GCLC	19.04	0.1281	0.0510	0.4470	0.1961	0.5864	0.4943	0.1936	1.19	1.31	1.33	1.51	1.21	2.10	-1.66
HSPA1A	18.91	0.0208	0.3329	0.8448	0.2111	0.5560	0.1404	0.3747	1.37	-1.17	1.12	-1.58	-2.09	1.46	2.47
TNF	18.22	0.0489	0.6451	0.2638	0.6774	0.8716	0.0660	0.4877	1.44	-1.11	-1.14	-1.23	-1.11	3.34	1.59
HMGB1	17.72	0.0630	0.0047	0.1891	0.2164	0.3968	0.6109	0.9320	1.24	1.50	1.39	1.48	1.05	1.16	-1.31
CYBB	17.34	0.0223	0.9489	0.3272	0.9352	0.2462	0.1217	0.5252	1.31	-1.01	1.05	1.03	-1.56	2.17	2.22
SERPINA1	17.27	0.0730	0.2471	0.9965	0.2167	0.5933	0.0938	0.5365	1.28	-1.21	1.06	-1.59	-2.06	1.66	2.28
MMP9	16.14	0.1858	0.1057	0.3289	0.1730	0.6537	0.5372	0.3199	1.36	-1.57	1.34	-2.39	-3.46	-1.31	2.56
CXCL1	15.91	0.1965	0.1305	0.5228	0.3722	0.4229	0.1961	0.4267	1.18	-1.26	1.09	-1.36	-1.94	1.46	2.02
EGR1	15.53	0.0263	0.3387	0.4572	0.2550	0.0909	0.6418	0.2217	1.33	-1.15	-1.04	1.48	-1.44	1.16	1.54
IL15	15.44	0.4922	0.0501	0.0862	0.3364	0.9222	0.3302	0.2744	1.10	1.37	-1.03	1.42	1.36	2.45	-1.15
DPP4	15.00	0.2114	0.4354	0.4450	0.1223	0.6509	0.1781	0.5079	1.15	-1.11	-1.14	-1.63	-1.95	1.18	-1.70
CD4	14.86	0.0816	0.4197	0.3359	0.9445	0.0935	0.4285	0.4223	1.19	-1.10	-1.11	1.02	-1.77	1.42	1.34
HMOX1	14.81	0.0587	0.8654	0.6194	0.2167	0.0801	0.9441	0.0954	1.27	-1.03	1.10	1.53	-1.44	1.59	3.05
PLA2G7	14.20	0.0865	0.9282	0.7051	0.4431	0.8269	0.0575	0.6794	1.27	1.02	1.17	-1.34	-1.49	2.31	3.07
PLAUR	13.46	0.2046	0.4947	0.9590	0.4582	0.2254	0.6593	0.1645	1.17	-1.10	1.07	-1.28	-2.15	-1.02	1.83
TIMP1	13.33	0.0658	0.5106	0.2000	0.4669	0.1399	0.8392	0.3714	1.23	-1.09	-1.17	1.25	-1.44	1.37	1.51
CD8A	13.23	0.0412	0.1604	0.8832	0.2939	0.7299	0.4394	0.8534	1.30	1.24	1.56	-1.44	-1.24	1.05	1.44
ADAM17	12.71	0.1919	0.8616	0.6441	0.1633	0.2687	0.7874	0.7724	1.15	1.02	1.29	1.49	-1.01	1.68	1.37
PTPRC	12.65	0.0279	0.7490	0.7072	0.5853	0.4100	0.2291	0.5759	1.24	1.04	1.20	-1.15	-1.53	1.41	1.53
PTGS2	12.55	0.0630	0.9846	0.6788	0.0812	0.9289	0.1552	0.9724	1.28	-1.00	1.15	-1.87	-1.79	1.17	1.18
IL1RN	12.41	0.3848	0.0842	0.1495	0.1451	0.6460	0.1061	0.8547	-1.13	-1.33	-1.03	-1.73	-1.38	1.48	2.18
ICAM1	11.99	0.2755	0.2793	0.7173	0.4375	0.4882	0.2839	0.3328	1.13	-1.16	1.06	-1.27	-1.68	1.31	2.10
APAF1	11.91	0.0852	0.8732	0.8466	0.8508	0.4636	0.0957	0.6159	1.20	-1.02	1.13	-1.05	-1.38	1.96	1.06
MNDA	11.64	0.0662	0.9441	0.3044	0.5092	0.6863	0.3526	0.3483	1.24	-1.01	-1.03	-1.23	-1.46	1.28	2.29
IL18BP	10.94	0.1220	0.0913	0.1308	0.6556	0.3362	0.4393	0.4857	1.14	1.18	1.06	1.11	-1.20	1.45	1.62
SERPING1	10.70	0.4339	0.2313	0.9868	0.5224	0.9825	0.7008	0.2993	-1.21	-1.41	-1.69	-1.52	-1.49	-1.03	5.36
IL1R1	10.49	0.1039	0.7632	0.9654	0.7602	0.3975	0.2078	0.9550	1.34	-1.07	1.28	-1.16	-2.00	2.25	1.40
IL1B	10.09	0.6746	0.3584	0.4113	0.3254	0.7338	0.1561	0.9919	-1.07	-1.19	1.01	-1.52	-1.84	1.68	1.40
TLR2	9.82	0.2787	0.6278	0.5153	0.7197	0.1332	0.2841	0.8863	1.15	1.08	1.06	1.14	-1.77	2.05	1.16
TNFSF5	9.48	0.3867	0.3648	0.1478	0.0599	0.4390	0.3545	0.6951	1.12	1.16	-1.12	-2.01	-1.39	-1.19	-1.18
IFI16	6.46	0.5860	0.6442	0.5085	0.2437	0.8924	0.4439	0.9634	1.09	1.09	-1.02	-1.62	-1.74	1.00	-1.02
CD19	5.24	0.1827	0.4304	0.1925	0.3513	0.1437	0.4404	0.5098	1.23	1.16	-1.03	1.48	-1.51	-1.11	-1.25

Values were computed only for white and Hispanic subjects for whom sex and age were recorded ($n = 68$). Unadjusted *P* values are shown for each effect, including interaction terms, and underlined (together with corresponding fold changes) when < 0.05. Fold changes for sex and ethnicity effects are computed by raising 2 to the power of the corresponding ΔC_T effect terms; for age effects, they are computed by multiplying the corresponding slope effect by the range of ages in the sample (69 - 23) and then exponentiating. HF indicates Hispanic female; HM, Hispanic male; WF, white female; WM, white male.

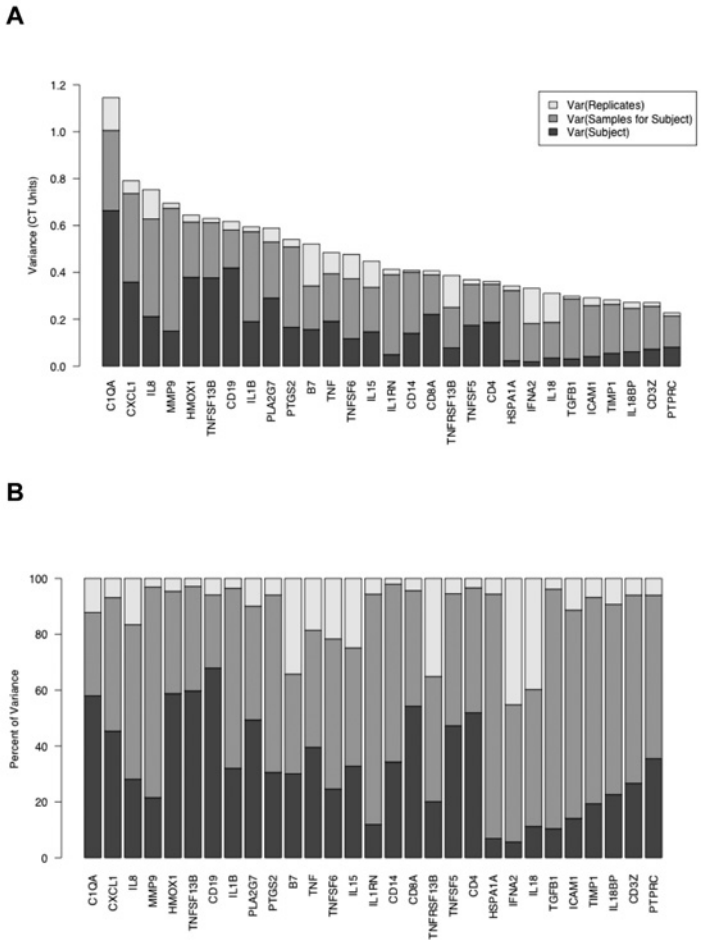


Figure 4. Source of variance in gene expression. (A) Variance components estimated from mixed-effect models, representing variation between subjects (dark grey), between longitudinal samples from same subject (grey), and between replicate RT-PCR reactions for same sample (white). Systematic variations affecting all samples drawn on same date have been subtracted before estimating variance components. (B) Variance components expressed as percentages relative to sum of components.

collected from healthy subjects injected with LPS. Healthy subjects who receive an injection of LPS experience mild fever and flu-like symptoms that subside within 24 h (35). Figure 5 shows the expression of a subset of genes with significant changes at any time point after LPS injection. Reference ranges (mean ± 2 SD) for healthy subjects are indicated by dashed lines. The plotted $\Delta\Delta C_T$ values are computed relative to the mean ΔC_T for the apparently healthy blood donors. Individual time courses are shown for each subject. Twenty-seven genes had significant changes in expression in LPS-injected subjects at any time postinfusion relative to apparently healthy blood donors, with adjusted false discovery rates of less than 5%. Each of these genes had pre-injection expression levels within the normal reference range for apparently healthy blood donors; each showed increased or decreased expression at 2 and/or 5 h postinfusion; and most returned to the normal expression range by 21 h after infusion. Fifteen genes increased or decreased expression by a factor greater than 10-fold, and 2 (MMP9 and IL1RN) increased more than 90-fold (Figure 5). Because the innate immune system's immediate response to LPS infusion is the production of inflammatory mediators by monocytes, it is not surprising that the genes showing substantial increases in expression include cytokines and chemokines associated with the monocyte/macrophage lineage, such as TNF, IL1B, CXCL1, and IL18. Key cell-surface markers (ICAM1, CD14) and signaling molecules (PTGS2/COX-2) also respond. Interestingly, the anti-inflammatory regulator IL1RN, which blocks the binding of IL1 to its receptor, was 1 of the 2 most overexpressed genes. This fits with the premise that inflammatory processes are tightly regulated by coordinated expression of pro-inflammatory and anti-inflammatory factors. These include genes with significant decreases in expression such as PLA2G7 and TNFSF5 (CD40 ligand) (see Figure 5).

DISCUSSION

The studies reported here are an initial step toward establishing normal reference ranges for the expression of genes related to inflammation and immunity. Several key observations emerged. First, the dynamic range of expression of most immune response genes is relatively limited among apparently healthy subjects. Second, expression levels for most genes analyzed are approximately log-normally distributed. Third, individuals exposed to bacterial endotoxin have gene expression profiles that are easily (albeit transiently) distinguished from those of an apparently healthy population. In developing the methods for these studies, it was also observed that multiple technical factors, including sample handling procedures, PCR reagents, and instrument calibration, contribute to the overall variation, which must be carefully controlled. Taken together, these observations support both the usefulness and practicality of establishing normal reference ranges for gene expression assays related to immune system function.

A variety of biological factors may contribute to the variation of expression observed in apparently healthy subjects (18). In general, these factors can be divided into intrinsic (for example, age, sex, genetics) and extrinsic (for example,

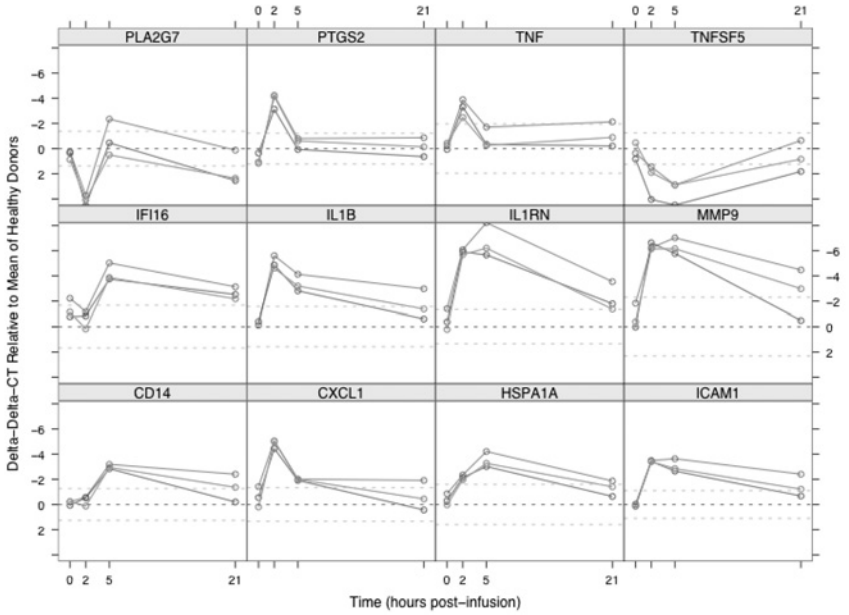


Figure 5. Time course of expression for 12 genes with significant responses to LPS infusion in 3 healthy male subjects. Whole blood was sampled at pre-LPS (0 h) and 2, 5 and 21 h post-LPS infusion. Gene expression is plotted as ΔC_T values relative to mean ΔC_T for healthy blood donors, with points and lines colored by subject. Mean and mean ± 2 SD are indicated by horizontal dashed lines. ΔC_T scale is inverted, so upward direction corresponds to increasing expression.

inflammatory, autoimmune disease, cancer, infections, and metabolism) factors. The apparently healthy blood donor population studied here may have included individuals with subacute illnesses or chronic conditions that contributed to the variability in expression of some immune response genes. Many chronic inflammatory and atopic diseases, such as arthritis, asthma, ulcers, gastritis, and allergies, are highly prevalent in the U.S. adult population, with frequencies ranging from 7% to 27% (36). Nonetheless, individuals with these conditions are deemed "healthy" and permitted to donate blood, provided these "chronic conditions are bring treated and the condition is under control," and they "feel well and are able to perform normal activities" (30).

Atherosclerosis is another highly prevalent condition which develops over several years and is asymptomatic in its early or even late stages. Several studies have demonstrated an elevation of C-reactive protein and other markers of inflammation in early stages of cardiovascular disease (37,38). Chronic infections with viruses (cytomegalovirus, Epstein-Barr virus, genital herpes, and human papillomavirus), bacteria (*Helicobacter pylori*), and protozoans (*Toxoplasma gondii*) also are common in the U.S. population, but do not consistently produce symptoms in immunocompetent persons. Periodic reactivation and suppression of these infections may account for some of the background variation in immune response gene expression. Dietary influences on immune system gene expression may include consumption of omega-3 fatty acids, arginine, and other nutrients as well as vegetarian diets (39,40).

Age, sex, and ethnicity also may contribute to the intersubject variation observed for several transcripts. However, the contributions of these factors appeared to be modest in the present study. Variations associated with age and sex have been previously reported (18,41,42), with some sex differences being directly attributable to differences in sex chromosomes (18). Several studies (18,42) have observed individual differences in interferon-responsive genes among individuals, suggesting further stratification in an apparently normal healthy subject group. Larger studies specifically targeting some of these factors are needed to elucidate the effects so that populations can be stratified for more precise diagnostic resolution.

Intrinsic and extrinsic factors can also alter the proportions of blood cell types such as neutrophils, monocytes, and lymphocytes, as well as the relative expression of individual transcripts within each cell type. These effects combine to produce the observed variation in transcript abundances in whole blood. The individual contributions of cell populations and gene regulation within cell types could be examined using flow cytometry combined with QRT-PCR, and deserve further study.

Given the variety of factors that can affect the expression of immune response genes in a blood donor population, it is remarkable that the overall dynamic range of expression is not wider than observed in the present study, whereas larger, up to 90-fold, but transient changes can be induced by the severe acute inflammatory stimulus LPS. In other diseases, such as rheumatoid arthritis and lupus, differences in gene expression from apparently healthy normals are more modest, 2- to 5-fold (43). These observations support the view that expression of these genes is maintained within limits by regulatory mechanisms, possibly to reduce the danger of tissue damage from constant activation of immune responses, while allowing appropriate responses to infectious threats. The limited dynamic range observed supports the development of expression-based diagnostics, allowing expression outside the normal reference range to indicate the presence of infections, cancer, or indolent autoimmune diseases.

Molecular diagnostics, including those based on gene expression, are increasingly being applied in the clinic (44,45). These tests have improved the selection of therapies, as well as dosage and treatment schedule. In addition, "treat-to-normal" strategies are routinely used in major

diseases such as hypertension and diabetes. Assays based on precise, quantitative measurements of immune system gene expression offer the promise of effective clinical monitors in infection, autoimmune diseases, and other immune-related conditions, such as transplant rejection and drug- or virus-induced immunosuppression, as well as cancer. A better understanding of the relevant factors that contribute to the individuality of gene expression in the human will help to establish the most appropriate normal reference values in the clinic and will serve as an essential step in the development of effective molecular diagnostics for these and other inflammatory and immunologic diseases.

ACKNOWLEDGMENTS

The authors would like to thank C. Edwards, C. Dinarello, A. Rasley, D. Nelson, M. Ascher, and C.T. Rigl for helpful comments, review, and discussion. This work was performed under the auspices of the Lawrence Livermore National Laboratory and was supported with funds from the Laboratory Directed Research and Development (LDRD) Program.

REFERENCES

1. Bild AH et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439:274-5.
2. Gladkevich A, Nelemans SA, Kauffman HF, Korf J. (2005) Microarray profiling of lymphocytes in internal diseases with an altered immune response: potential and methodology. *Mediators Inflamm.* 2005:317-30.
3. Han D, Leith J, Alejandro R, Bolton W, Ricordi C, Kenyon NS. (2005) Peripheral blood cytotoxic lymphocyte gene transcript levels differ in patients with long-term type 1 diabetes compared to normal controls. *Cell Transplant.* 14:403-9.
4. Perez EA, Pusztai L, Van de Vijver M. (2004) Improving patient care through molecular diagnostics. *Semin. Oncol.* 31(5 Suppl 10):14-20.
5. Baechler EC et al. (2003) Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl. Acad. Sci. U. S. A.* 100:2610-5.
6. Rus V, Chen H, Zernetkina V, Magder LS, Mathai S, Hochberg MC, Via CS. (2004) Gene expression profiling in peripheral blood mononuclear cells from lupus patients with active and inactive disease. *Clin. Immunol.* 112:231-4.
7. Chen X et al. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell* 13:1929-39.
8. Gang J et al. (2005) Discovery and analysis of pancreatic adenocarcinoma genes using DNA microarrays. *World J. Gastroenterol.* 11:6543-8.
9. Chang JC et al. (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* 362:362-9.
10. US patent no. 6,960,439: Identification, monitoring and treatment of disease and characterization of biological condition using gene expression profiles, covering the use of a Healthy Normals Reference Dataset, issued to Source MDx, Nov. 4, 2005.
11. Stitt JT (1979) Fever versus hyperthermia. *Fed. Proc.* 38:39-43
12. Conti B, Tabarean I, Andrei C, Bartfai T. (2004) Cytokines and fever. *Front. Biosci.* 9:1433-49.
13. Dinarello CA. (2004) Infection, fever, and exogenous and endogenous pyrogens: some concepts have changed. *J. Endotoxin Res.* 10:201-22.
14. Jiang H, Chess L (2004) An integrated view of suppressor T cell subsets in immunoregulation. *J. Clin. Invest.* 114:1198-1208.
15. Campbell C, Vernon SD, Karem KL, Nisenbaum R, Unger ER. (2002) Assessment of normal variability in peripheral blood gene expression. *Dis. Markers* 18:201-6.
16. Chowers I, Liu D, Farkas RH, et al. (2003) Gene expression variation in the adult human retina. *Hum. Mol. Genet.* 12:2881-93.
17. Cole J, Tsou R, Wallace K, Gibran N, Isik F. (2001) Comparison of normal human skin gene expression using cDNA microarrays. *Wound Repair Regen.* 9:77-85.
18. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO. (2003) Individuality and variation in gene expression patterns in whole blood. *Proc. Natl. Acad. Sci. U. S. A.* 100:1896-1901.
19. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33:422-5
20. Baechler EC, Batliwalla FM, Karypis G, et al. (2004) Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation. *Genes Immun.* 5:347-53.
21. Debey S, Schoenbeck U, Hellmich M, Gathof BS, Pillai R, Zander T, Schultze JL. (2004) Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. *Pharmacogenomics J* 4:193-207.
22. Han ES, Wu Y, McCarter R, Nelson JF, Richardson A, Hilsenbeck SG (2004) Reproducibility, sources of variability, pooling and sample size: important considerations for the design of high-density oligonucleotide array experiments. *J. Gerontol. A Biol. Sci. Med. Sci.* 59:306-15.
23. Rainen L, Oelmueller U, Jurgensen S, et al. (2002) Stabilization of mRNA expression in whole blood samples. *Clin. Chem.* 48:1883-90.
24. Tan PK, Downey TJ, Spitznagel EL, et al. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* 31:5676-84.
25. Snider JV, Wechser MA, Lossos IS. (2001) Human disease characterization: real-time quantitative PCR analysis of gene expression. *Drug Discov. Today* 6:1062-7.
26. Liles WC, Van Voorhis WC. (1995). Nomenclature and biological significance of cytokines involved in inflammation and the host immune response. *J. Infect. Dis.* 172:1573-80.
27 Joyce DA, Steer JH, Beilharz MW, Stranger R. (1995). A system for assessment of monokine gene expression using human whole blood. *Genet. Anal.* 12:39-43.
28. Inoue M, Nishimura S, Hori G, Nakahara H, Saitom M, Yoshihara Y, Amari S. (2004) Improved parameter estimation for variance-stabilizing transformation of gene-expression microarray data. *J. Bioinform. Comput. Biol.* 2:669-79.
29. Naef F, Hacker CR, Patil N, Magnasco M. (2002) Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol.* 3(4) Epub 2002 Mar 22
30. American Red Cross: www.redcross.org
31. Heid CA, Stevens J, Livak KJ, Williams PM (1996) Real time quantitative PCR. *Genome Res.* 6:986-94.
32. Livak KJ, Schmittgen TD. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-??CT}$ method. *Methods* 25:402-8.
33. Pinheiro J, Bates DM. (2000) *Mixed-Effects Models in S and S-PLUS*. New York: Springer. 528 pp
34. R Development Core Team (2004) R Foundation for Statistical Computing, Vienna, Austria.
35. Martich G, Boujoukos A, Suffredini A (1993) Response of man to endotoxin. *Immunobiology* 187:403-16.
36. Schiller JS, Adams PF, Nelson ZC (2005) Summary health statistics for the US population: National health interview survey, 2003. *Vital Health Stat. 10* Apr(224):1-104.
37. Koenig W et al. (1999) C-Reactive protein, a sensitive marker of inflammation, predicts future risk of coronary heart disease in initially healthy middle-aged men: results from the MONICA (Monitoring Trends and Determinants in Cardiovascular Disease) Augsburg Cohort Study, 1984 to 1992. *Circulation* 99:237-42.
38. Pearson TA et al. (2003) Markers of inflammation and cardiovascular disease: application to clinical and public health practice: A statement for healthcare professionals from the Centers for Disease Control and Prevention and the American Heart Association. *Circulation* 107:499-511.
39. Bistrian BR. (2004) Practical recommendations for immune-enhancing diets. *J. Nutr.* 134:2868S-72S.
40. Simopoulos AP. (2002) Omega-3 fatty acids in inflammation and autoimmune diseases. *J. Am. Coll. Nutr.* 21:495-505.
41. Eady J et al. (2005) Variation on gene expression profiles of peripheral blood mononuclear cells from healthy volunteers. *Physiol. Genomics* 22:402-11.
42. Radich J et al. (2004) Individual-specific variation of gene expression in peripheral blood leukocytes. *Genomics* 83:980-8.
43. Tryon V et al. High-precision gene expression analysis of rheumatoid arthritis and other inflammatory diseases. *Int. Assoc. Inflammation Soc.* Meeting Poster Presentation, Vancouver BC, August 2003.
44. Ross J et al. (2003) The HER-2/neu gene and protein in breast cancer 2003: biomarker and target of therapy. *Oncologist* 8:307-25.
45. Madhusudan S, Ganesan TS (2004) Tyrosine kinase inhibitors in cancer therapy. *Clin. Biochem.* 37:618-35.

Aim 3

Differentiate the host response surrogate marker profile of infected sick or presyndromic animals or humans (year 2 -3).

Original Report: Laboratory Investigation

American Journal of Nephrology

Am J Nephrol 2004;24:268–274
DOI: 10.1159/000077409

Received: November 25, 2003
Accepted: February 11, 2004
Published online: March 19, 2004

Serum Protein Profile Alterations in Hemodialysis Patients

Richard G. Langlois[a] James E. Trebes[a] Enrique A. Dalmasso[b] Yong Ying[b]
Robert W. Davies[c] Mario P. Curzi[c] Bill W. Colston, Jr.[a]
Kenneth W. Turteltaub[a] Julie Perkins[a] Brett A. Chromy[a] Megan W. Choi[a]
Gloria A. Murphy[a] J. Patrick Fitch[a] Sandra L. McCutchen-Maloney[a]

[a]Lawrence Livermore National Laboratory, Livermore, Calif.; [b]Ciphergen Biosystems, Fremont, Calif. and [c]Diablo Nephrology Medical Group, Inc., Walnut Creek, Calif., USA

Key Words
Hemodialysis · SELDI-TOF-MS · Protein profiling · Proteomics · Biomarkers

Abstract
Background: Serum protein profiling patterns can reflect the pathological state of a patient and therefore may be useful for clinical diagnostics. Here, we present results from a pilot study of proteomic expression patterns in hemodialysis patients designed to evaluate the range of serum proteomic alterations in this population. ***Methods:*** Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) was used to analyze serum obtained from patients on periodic hemodialysis treatment and healthy controls. Serum samples from patients and controls were first fractionated into six eluants on a strong anion exchange column, followed by application to four array chemistries representing cation exchange, anion exchange, metal affinity and hydrophobic surfaces. A total of 144 SELDI-TOF-MS spectra were obtained from each serum sample. ***Results:*** The overall profiles of the patient and control samples were consistent and reproducible. However, 30 well-defined protein differences were observed; 15 proteins were elevated and 15 were decreased in patients compared to controls. Serum from 1 patient exhibited novel protein peaks suggesting possible additional changes due to a secondary disease process. ***Conclusion:*** SELDI-TOF-MS demonstrated consistent serum protein profile differences between patients and controls. Similarity in protein profiles among dialysis patients suggests that patient physiological responses to end-stage renal disease and/or dialysis therapy have a major effect on serum protein profiles.

Introduction

Proteomics can be defined as the characterization of total protein composition of an organism [1]. Comparative proteomic analysis under different physiological states may be a powerful approach for identifying biomarkers of health status, since many proteins that are secreted into bodily fluids are differentially expressed in

KARGER
Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0250–8095/04/0242–0268$21.00/0

Accessible online at:
www.karger.com/ajn

Sandra L. McCutchen-Maloney, PhD
Biodefense Division, Lawrence Livermore National Laboratory
7000 East Ave, L-452
Livermore, CA 94550 (USA)
Tel. +1 925 4235065, Fax +1 4222282, E-Mail smaloney@llnl.gov

response to physiological changes such as infection or inflammation. Identification of proteins characteristic of a specific disease may provide biomarkers that can be used in simple, non-invasive clinical diagnostics [2–4].

One approach to identify differentially expressed proteins is surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS). SELDI-TOF-MS is an array-based MS technology introduced by Hutchens and Yip [5] that utilizes selective adsorption of a subset of proteins in a given sample to array surfaces differing in chemical coating [6]. Arrays are functionalized for ion exchange, immobilized metal affinity, or hydrophobic selectivity allowing the serum sample to be fractionated into subsets of proteins with similar chemical affinity. Proteins captured on the array are ionized, and their masses are determined by time-of-flight (TOF) MS. A principle advantage of SELDI-TOF-MS is the ability to rapidly screen hundreds to thousands of proteins for differences between diseased individuals and control subjects, even if the protein functions and identities are unknown. Thus, this technique provides a broad unbiased screen for protein expression differences. Once a candidate protein is detected, however, additional experimental work is required to determine the identity and function of the candidate biomarker.

To date, the SELDI-TOF-MS technique has primarily been used to screen for candidate biomarkers for specific diseases. This approach has yielded potential biomarkers for prostate, bladder, lung, breast and ovarian cancers as well as Alzheimer's disease [7–12]. In addition, we believe that this approach has considerable potential for monitoring patients with complex chronic conditions or syndromes to identify episodes of relapse, infection, or drug failure. There is one report, for example, of the analysis of urine protein profiles to characterize renal allograft rejection [13]. Analysis of patients with chronic conditions, however, is complicated by protein profile alterations due to the underlying condition and potential patient-to-patient variability in disease state. The ability to rapidly screen large numbers of protein types per patient provides a detailed protein profile facilitating interpretation of these complex factors [14, 15]. Consequently, the present pilot study was designed to compare serum samples from hemodialysis patients with samples from healthy controls to investigate the effects of end-stage renal disease on serum protein profiles. In the future, it is hoped that protein profiles may help to identify infections or other complications in dialysis patients [16].

Before SELDI-TOF-MS can be applied to studies of complications in dialysis patients, it is important to understand the effects of end-stage renal disease and dialysis treatments on serum protein profiles. Kidney failure can be caused by a variety of underlying complications including diabetes, hypertension, and glomerulonephritis, and each of these etiologies could have a different effect on serum components. The dialysis process itself alters the concentrations of low- vs. high-molecular-weight proteins in serum depending on the time of sampling. Protein profiles could also be altered by patient responses to the hemodialysis process (e.g. inflammation, cytokine production). Finally, patient-to-patient variation in the presence of other chronic diseases or health complications may be important. While there is a growing literature characterizing specific serum proteins and metabolites in hemodialysis patients [17–20], the focus of this study is to begin to evaluate a broad profile of serum proteins in patients vs. control individuals in order to understand the effects of the complexities described above. A better understanding of these issues would facilitate future application of protein profiles to the diagnosis of complications in dialysis patients.

Materials and Methods

Protocols for this study were reviewed and approved by the LLNL Institutional Review Board and comply with NIH guidelines. Blood samples were obtained with informed consent from 4 unaffected healthy control subjects, and 4 patients that are receiving dialysis treatments three times per week as a consequence of renal failure. Samples from dialysis patients were obtained prior to their routine dialysis session. The 4 dialysis patients (subjects 1–4) consisted of 3 females and 1 male between the ages of 29 and 63 years. Causes of renal failure differed for each of these 4 patients. End-stage renal disease was secondary to the following causes: diabetes, cyclosporine toxicity, IgA nephropathy, and hypertension. The 4 control subjects (subjects 5–8) consisted of 2 females and 2 males, with an age range of 32–52 years. Blood from all subjects were collected in 2.5 ml BD vacutainer SST glass serum tubes (Becton Dickinson, Franklin Lakes, N.J., USA) and spun at 2,500 rpm at 4°C for 30 min. The separated serum was divided into 0.1-ml aliquots and stored at –80°C until analysis. All samples were coded before sample preparation and MS analysis. SELDI-TOF-MS analysis was performed blindly with no knowledge of the source of the samples. After the experimental work was completed, results were identified as coming from patient or control group samples to compare protein profiles between groups.

Frozen serum samples were prepared for SELDI-TOF-MS as outlined in figure 1. Each serum sample (subjects 1–8) was thawed and spun at 20,000 *g* for 10 min at 4°C. 30 µl of pH 9.0 buffer (9 *M* urea/2% CHAPS/50 m*M* Tris-HCl) was added to 20 µl of each serum sample before mixing with Q Ceramic HyperD® F beads (Ciphergen Biosystems, Fremont, Calif., USA) in a filtration plate. Proteins were eluted through the filter by washes with buffers of different pH. Fraction 1 (F1) consisted of flow through and material eluted with 200 µl

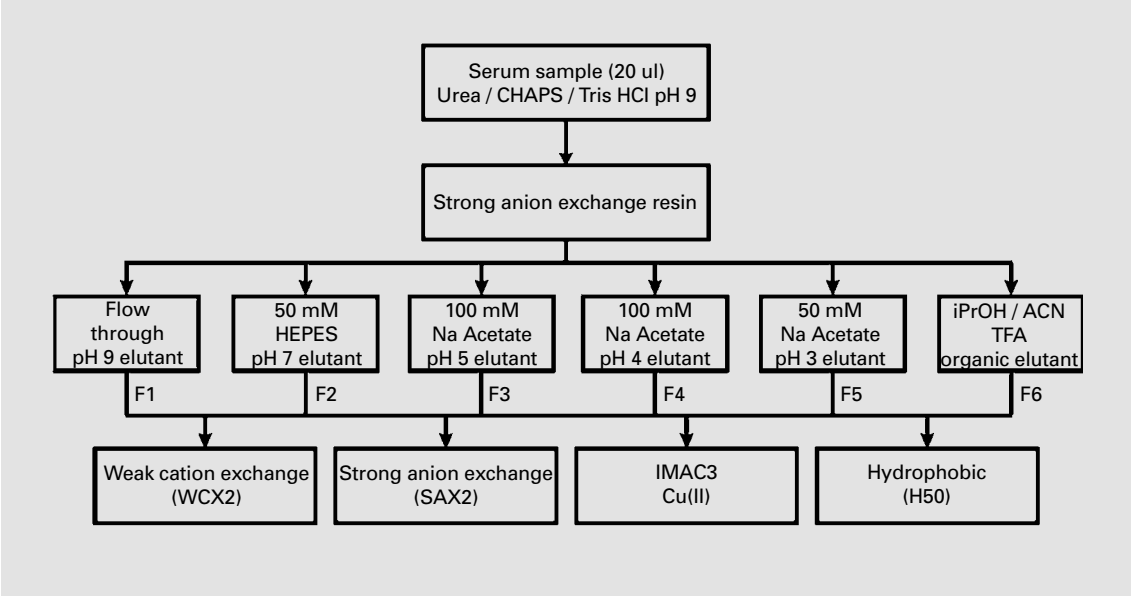


Fig. 1. Flow chart of serum processing for SELDI-TOF-MS: Elution from a strong ion exchange resin with a pH gradient to yield six fractions followed by application of each fraction to four different ProteinChip® Array surfaces.

of pH 9.0 buffer (50 m*M* Tris-HCl, 0.1% *n*-octyl glucopyranoside, OPG). This procedure was repeated for pH 7 buffer (50 m*M* Hepes, 0.1% OPG), pH 5 buffer (100 m*M* sodium acetate, 0.1% OPG), pH 4 buffer (100 m*M* sodium acetate, 0.1% OPG), pH 3 buffer (50 m*M* sodium acetate, 0.1% OPG) and an organic solvent buffer (33.3% isopropanol/16.7% acetonitrile/0.1% trifluoroacetic acid) to give fractions 2 through 6 (F2–F6) respectively. Each fraction was then applied onto four different Ciphergen ProteinChip® Arrays: Weak Cation exchange (WAX2), Strong Anion exchange (SAX2), Immobilized Metal Affinity Capture (IMAC) (Copper II), and Hydrophobic (H50) surfaces using 96-well ProteinChip Array BioProcessors. Each array surface was prepared using standard protocols described in the Ciphergen ProteinChip® Applications guide [21]. The energy-absorbing molecules (EAMs), α-cyano-4-hydroxy cinnamic acid (CHCA) and sinapinic acid (SPA) were deposited on the array spots and allowed to air dry. Different EAMs and laser powers were used to optimize detection for proteins differing in molecular weight (MW). CHCA was used as the EAM for proteins with a MW < 15 kDa, while SPA was used primarily for proteins with MW > 15 kDa. These fractionations provide a broad coverage of proteins based on chemical class rather than function. A total of 144 TOF mass spectra analyzing proteins with mass to charge ratio (m/z) from 1 to 200 kDa were obtained for each sample (reflecting 72 different conditions in duplicate). For SELDI-TOF-MS, proteins and peptides were detected using a Ciphergen PBS-IIC ProteinChip® Reader, a time-lag focusing, linear, laser desorption/ionization TOF-MS. All spectra were acquired in the positive-ion mode. Each spectrum was an average of 130 laser shots and externally calibrated against a mixture of known peptides or proteins. The spectra were analyzed using the Biomarker Wizard function in ProteinChip® Software v3.1.1.

Results and Discussion

Overall, the 8 serum samples yielded qualitatively similar protein profiles with the 72 different fractionation and ProteinChip® Array conditions. The data in figure 2a show a typical example with the major peaks very consistent among all dialysis patients and all controls, with a few minor peaks varying between individuals. It is difficult to quantify the total number of protein features analyzed from each sample because some features appear in multiple array conditions, and some minor features are hard to differentiate from noise. Experience with previous studies and literature reports provide an estimate that about 500–1,000 protein features per sample are detected in a study of this size [15].

A number of clearly defined peaks were observed that consistently distinguish the patient samples (1–4) from the control samples (5–8) across the 72 analysis conditions. Two spectra were chosen to illustrate differences between patients and controls. The spectra in figure 2b show peaks at 5.8 and 11.7 kDa that have greater intensity in all patients compared with controls, while peaks at 7.7 and 9.3 kDa have reduced intensity in patients compared with controls. A close-up view from another fraction and EAM shows two of these peaks at 9.3 and 11.7 kDa that consistently distinguish patients from controls (fig. 2c).

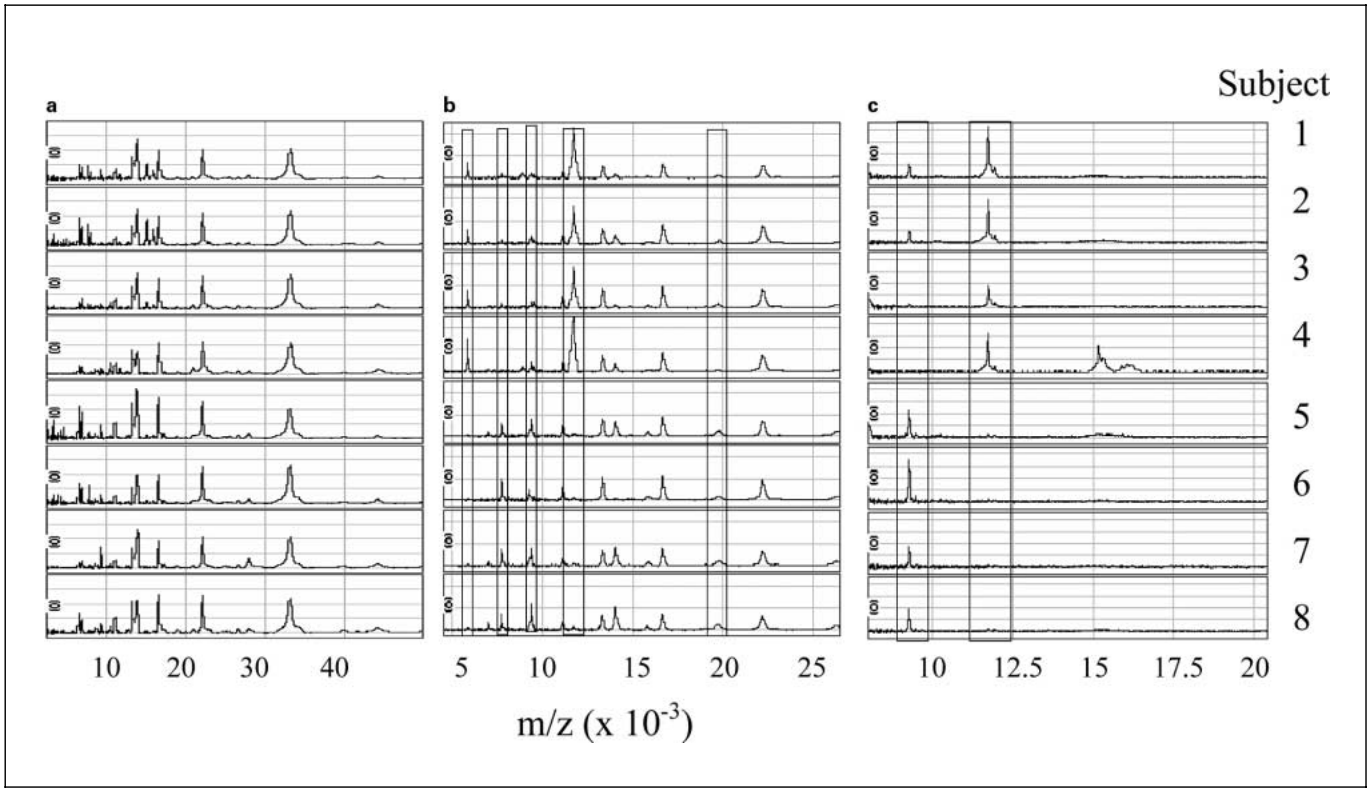


Fig. 2. Mass spectra of serum samples from the 8 subjects in the study. Subjects 1–4 are dialysis patients, while subjects 5–8 are unaffected healthy controls. **a** Fraction 4, WCX array surface with CHCA. Note similar peaks with all samples. **b** Fraction 5, IMAC array surface with CHCA. Note candidate protein markers at 5.8, 7.7, 9.3, 11.7 and 19.7 kDa. **c** Fraction 2, IMAC array surface with SPA and high laser power. Note candidate protein markers at 9.3 and 11.7 kDa, and the unique peaks for subject 4 at 15.5 and 15.9 kDa.

The majority of peaks, however, show similar amplitudes among all samples.

A listing of protein peaks that differ between patients and controls is shown in table 1. A total of 15 candidate proteins showed increased intensity in at least 3 out of 4 patients compared with all controls, while 15 candidates showed decreased intensity in at least 3 out of 4 patients compared with all controls. For 60% of these candidate protein peaks, intensities for all 4 patients were outside the range for all 4 controls. Thus, most candidate proteins clearly distinguish all patients from all controls in this study. A few samples are listed as outliers in table 1 as they lacked, or in some cases contained, one or more peaks that were characteristic of their group. In addition, data from subject 4 showed two strong peaks at 15.2 and 15.9 kDa that were not present in any of the other 7 samples (fig. 2c), suggesting that another factor besides dialysis may be responsible for these peaks.

Unfortunately, substantial additional biochemical analysis is required to determine the identity of each of these 30 candidate proteins. It may ultimately be possible to correlate the profile of protein differences with disease status without peak identification. Our hope, however, is that with further studies of patients and controls, we can focus on a smaller number of diagnostic peaks for identification that may contribute to a better understanding of end-stage renal disease.

The results of this SELDI-TOF-MS study provide an overview of serum protein profile alterations in hemodialysis patients. While it is difficult to quantify the exact frequency of protein alterations, our observation of 30 candidate protein biomarkers that distinguish the two populations is much larger than the 1–5 candidate markers reported from similar studies on specific diseases [9, 12, 13, 22, 23]. Thus, dialysis treatment, or clinical factors present in end-stage renal disease, have a dramatic effect

Table 1. Candidate protein peaks that distinguish healthy controls from hemodialysis patients

MW kDa	Peak height in patients vs. control	Fraction	Chip surface	EAM/laser intensity	Outlier sample #
78.8	Lower	2	H50	SPA high	6
51.3	Higher	5	H50	SPA high	
50.8	Higher	4	IMAC	SPA high	
45.3	Lower	4	H50	SPA high	6
43.4	Higher	4, 6	H50	SPA high	
25.5	Higher	5	IMAC	SPA high	
20.9	Higher	5	WCX	SPA high	
19.7	Lower	3	IMAC	CHCA	
17.3	Lower	6	SAX	SPA low	
15.9	Higher	4, 6	H50, IMAC, WCX	SPA low	4
15.2	Higher	4, 6	H50, IMAC, WCX	SPA high, low	4
14.7	Higher	1	IMAC	SPA high	
14.1	Lower	6	SAX	SPA low	
13.3/13.4	Higher	1	H50, IMAC	All	
12.8	Lower	1	H50	SPA low	
12.6	Lower	5	SAX	SPA high	
12.1	Lower	5	SAX	SPA high	
11.7	Higher	2, 3	H50, IMAC	All	
10.3	Lower	1	IMAC	SPA high	8
9.3	Lower	1, 2, 3	H50, IMAC	All	
8.6	Higher	1	H50	SPA low, CHCA	
8.6	Lower	6	SAX	SPA low, CHCA	
8.2	Lower	5	H50	SPA low	
7.7/7.8	Lower	3, 4, 6	IMAC, WCX	SPA low, CHCA	
7.1	Higher	6	WCX	SPA low	
6.4	Lower	5	H50	SPA low	
5.8	Higher	3	IMAC	SPA low, CHCA	
4.3	Higher	1	H50	CHCA	4
2.7	Lower	1	WCX	CHCA	
1.9	Higher	1	WCX	CHCA	

Note: Fraction, Chip surface, and EAM/laser intensity indicate the experimental conditions used where the candidate peak was observed. Multiple entries (e.g. 9.3 kDa) indicate that the candidate peak was observed using several experimental conditions.

on serum protein profiles. The 4 dialysis patients share most of these 30 protein alterations, and more than half of the marker changes are shared by all patients compared with all controls. This suggests that renal failure in general, or dialysis therapy, both of which are shared by all patients, may have a greater effect on protein profile alterations than the underlying causes of kidney failure that differed among all 4 patients.

A dialysis treatment effect could result from either differential loss of low MW components through the dialysis membrane, or from patient responses to dialysis such as the production of cytokines or inflammatory response proteins. The data in table 1 show that biomarker proteins vary in MW from 1.9 to 78.8 kDa, and that the biomarkers elevated in patients were spread across the full MW range. This suggests that patient physiological responses to dialysis are more important than dialysis membrane fractionation in producing the observed protein profile patterns. Finally, the unique protein markers observed in patient 4 suggest other clinical factors may be present in this individual in addition to end-stage kidney disease. One clinical factor that is unique to patient 4 is that this is the only subject with hepatitis C. Further studies would be required to determine if hepatitis or liver damage are the cause of differential protein markers seen in this patient.

In summary, SELDI-TOF-MS provides a convenient, rapid method for screening large numbers of serum proteins to characterize protein profile alterations in complex clinical conditions. This pilot study was designed to provide insights into the effects of end-stage renal disease and dialysis treatments on serum protein profiles. Our results show that although a number of factors in hemodialysis patients such as secondary diseases must be considered, SELDI-TOF-MS may be useful in the future as a diagnostic tool to identify treatment complications and potentially reduce patient mortality. Our results show that while patients differ dramatically from controls, the protein profiles of dialysis patients are similar to each other. This suggests that there may be characteristic profile for dialysis patients. The unique features in patient 4 support the potential of detecting additional clinical conditions. Future studies with larger numbers of dialysis patients and control individuals will be required to determine whether treatment-related complications could also be detected using this approach.

Acknowledgments

This work was performed under the auspices of the US Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48, with support from Laboratory Directed Research and Development funding. UCRL-JRNL-201081.

References

1 Tyers M, Mann M: From genomics to proteomics. Nature 2003;422:193–197.
2 Hanash S: Disease proteomics. Nature 2003; 422:226–232.
3 Anderson NL, Anderson NG: The human plasma proteome. Mol Cell Proteomics 2002;1: 845–867.
4 Kennedy S: Proteomic profiling from human samples: The body fluid alternative. 2001;120: 379–384.
5 Hutchens TW, Yip TT: New desorption strategies for the mass spectrometric analysis of macromolecules. Rapid Commun Mass Spectrom 1993;7:576–580.
6 Issaq HJ, Conrads TP, Prieto AP, Tirumalai R, Veenstra TD: SELDI-TOF-MS for diagnostic proteomics. Anal Chem 2003;149A–155A.
7 Paweletz CP, Liotta LA, Petricoin EF: New technologies for biomarker analysis of prostate cancer progression: Laser capture microdissection and tissue proteomics. Urology 2001;57: 160–163.
8 Vlahou A, Schellhammer PF, Medrinos S, Patel K, Kondylis FI, Gong L, Nasim S, Wright GL Jr: Development of a novel approach for the detection of transitional cell carcinoma of the bladder in urine. Am J Pathol 2001;158: 1491–1501.
9 Zhukov TA, Johanson RA, Cantor AB, Clark RA, Tockman MS: Discovery of distinct protein profiles specific for lung tumors and premalignant lung lesions by SELDI mass spectrometry. Lung Cancer 2003;40:267–279.
10 Wulfkuhle JD, McLean KC, Pawaletz CP, Sgroi DC, Trock BJ, Steeg PS, Petricoin EF III: New approaches to proteomic analysis of breast cancer. Proteomics 2001;1:1205–1215.
11 Petricoin EF III, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA: Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–577.
12 Carrette O, Demalte I, Scherl A, Yalkinoglu O, Corthals G, Burkhard P, Hochstrasser DF, Sanchez JC: A panel of cerebrospinal fluid potential biomarkers for the diagnosis of Alzheimer's disease. Proteomics 2003;3:1486–1494.
13 Clark W, Silverman BC, Zhang Z, Chan DW, Klein AS, Molmenti EP: Characterization of renal allograft rejection by urinary proteomic analysis. Ann Surg 2003;237:660–665.
14 Issaq HJ, Veenstra TD, Conrads TP, Felschow D: The SELDI-TOF-MS approach to proteomics: Protein profiling and biomarker identification. Biochem Biophys Res Commun 2002; 292:587–592.
15 Poon TCW, Yip TT, Chan ATC, Yip C, Yip V, Mok TSK, Lee CCY, Leung TWT, Ho SKW, Johnson PJ: Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. Clin Chem 2003;49:752–760.

16 Allon M, Depner TA, Radeva M, Bailey J, Beddhu S, Butterly D, Coyne DW, Gassman JJ, Kaufman AM, Kaysen GA, Lewis JA, Schwab SJ; HEMO Study Group: Impact of dialysis dose and membrane on infection-related hospitalization and death: results of the HEMO Study. J Am Soc Nephrol 2003;14:1863–1870.
17 Kaysen GA, Durbin JA, Muller HG, Mitch WE, Rosales L, Levin NW; HEMO Group: Impact of albumin synthesis rate and the acute phase response in the dual regulation of fibrinogen levels in hemodialysis patients. Kidney Int 2003;63:315–322.
18 Caglar K, Peng Y, Pupim LB, Flakoll PJ, Levenhagen D, Hakim RM, Ikizler TA: Inflammatory signals associated with hemodialysis. Kidney Int 2002;62:1408–1416.
19 Tarakcioglu M, Erbagci AB, Usalan C, Deveci R, Kocabas R: Acute effect of hemodialysis on serum levels of the proinflammatory cytokines: Mediators Inflamm 2003;12:15–19.
20 Kalantar-Zadeh K, Don BR, Rodriguez RA, Humphreys MH: Serum ferritin is a marker of morbidity and mortality in hemodialysis patients. Am J Kidney Dis 2001;37:564–572.
21 Ciphergen ProteinChip® Applications guide 107–118. Fresno/CA, Ciphergen Inc, USA.
22 Shiwa M, Nishimura Y, Wakatabe R, Fukawa A, Arikuni H, Ota H, Kato Y, Yamori T: Rapid discovery and identification of a tissue-specific tumor biomarker from 39 human cancer cell lines using the SELDI ProteinChip platform. Biochem Biophys Res Commun 2003;309:18–25.
23 Ye B, Cramer DW, Skates SJ, Gygi SP, Pratomo V, Fu L, Horick NK, Licklider LJ, Schor JO, Berkowitz RS, Mok SC: Haptoglobin-α subunit as potential serum biomarker in ovarian cancer: Identification and characterization using proteomic profiling and mass spectrometry. Clin Cancer Res 2003;9:2904–2911.

Early detection of infectious disease using host biochemical signatures in mice infected with cowpox virus

Richard G. Langlois, Lorena Diehl, Kodumudi S. Venkateswaran, Kevin S. McLoughlin, Michael S. Ascher, Fred P. Milanovich, C. Rick Lyons, and Kenneth W. Turteltaub

Abstract

Introduction

There is growing urgency to develop techniques for rapid detection and diagnosis of infectious disease in human populations. Rapid detection is critical for reducing the morbidity and mortality from either bio-terrorism events or newly emerging diseases. Current methods for direct agent detection using culture methods or microbial component detection using antibodies or PCR have a number of limitations. Rapid microbial detection in blood may not be possible for agents that remain localized to the site of infection or agents that do not appear in peripheral blood until the later stages of the disease. Cell culture and sample enrichment procedures can also require several days. Finally, newly emerging diseases or genetically modified organisms may have never been seen before complicating organism-specific detection methods.

Host responses may provide early signals in blood even from localized infections. Cells in the innate immune system produce a rapid response after initial contact with a potential pathogen. While pathogen responses initially involve local cell signaling processes designed to activate near-by immune cells, cascades of cytokines and chemokines are released into the periphery to activate additional cells types and to cause them to migrate to the site of the infection. Thus, early immune responses may provide general indicators for the presence of many different infection types. A spectrum of innate and adaptive immune markers, in combination with other biochemical markers may be required to obtain more disease-specific detection.

Direct studies of the time course of natural diseases in humans complicated by the difficulty in defining exposure doses and exposure timing. Model systems, both in cell culture and animal models, allow precise control over exposure dose and timing. Studies of specific tissue types in culture or in whole animals have been used to define early responses in cells at the site of infection, but it is often difficult to relate these tissue-specific responses to systemic responses in the whole animal. Population screening for early disease detection is facilitated by the use off minimally invasive sampling techniques involving fluids such as blood, urine, or saliva. Serum samples were selected for these studies because serum is easily collected and stored; serum also contains a wide variety of proteins with well-regulated concentrations in healthy animals. While serum has the advantage of integrating systemic effects of localized infections in different organs, some protein markers may not be detected because they do not leave the site of infection, or proteins produced by normal tissues dilute their concentration changes.

Cowpox infection in mice was chosen as the model system for these studies. Mice have been shown to be susceptible to infection by both cowpox and vaccinia viruses [5,6]. The severity of infection varies from mild to lethal depending on the strain of mouse, strain of virus, and the location and dose of viral challenge. Viral instillation in the lung of mice produces a pulmonary infection that has been used as a model for pulmonary smallpox infection in humans, so that this model has been used extensively for studies of anti-viral drugs [18,19].

The TK- strain of cowpox virus into BALB/c mice model used for these studies exhibits three major features important to early detection of infections. First, there is an incubation period of about 6 days before the mice show signs of illness. This provides the opportunity to assess protein changes in serum throughout this prodromal period, as well as the period of active infection. Second, this model produces a localized infection in the lung, with no live virus detectable in the blood using plate assays. Thus, this model is well suited to analyze whole-animal systemic responses in blood to a localized infection. Finally, the conditions used for these studies produce a serious illness, but no lethality was observed from viral infection. This feature allows assessment of the early stages of recovery from infection, and insures that the biochemical changes we observe reflect moderate disease rather than the severe toxicity of super-lethal doses.

A variety of approaches have been used to select host biochemical markers that may be responsive to infectious diseases. Early studies focused on detailed studies of one, or a few, protein markers known to be associated with specific diseases. More recently, a number of broad screening methods have been developed as discovery tools to search for new markers among protein or mRNA targets. Commercial and custom gene expression assays allow the screening of greater than 10,000 gene targets per sample [7,8]. Protein screening methods including liquid chromatography mass spectrometry (LCMS) [13], surface enhanced laser desorption ionization mass spectrometry (SELDI-MS) [14], and two-dimension gel electrophoresis are also being evaluated as discovery tools for protein markers [15,20,21]. While these approaches provide powerful tools for screening thousands of gene products, they have several limitations for rapid disease detection. One limitation of discovery methods is that sample preparation, data collection, and data analysis frequently take 24 hours or more before an analysis is completed. A second limitation is that both chip hybridization and mass spectrometric results are difficult to quantify, particularly for samples with a wide range of analyte concentrations. Finally, these techniques may not have the dynamic range to quantify high-abundance proteins (e.g. Apo-A1 at about 0.1 mg/ml), and low abundance signaling proteins (e.g. Interlukin-3 at about 10 pg/ml).

The technique of multiplex immunoassay of serum proteins was selected for the present study. Antibodies have been developed for a variety of serum proteins, including proteins involved in inflammation, immunity, cell signaling, cell proliferation, lipid transport, etc. [24,25] Immunoassays have been developed for many of these protein types, with well-characterized commercially available kits for the analysis of multiple proteins targets [26,27], and companies that provide fee for service analyses of samples with specific analyte panels [Rules Based Medicine Inc., Austin TX; Upstate Cell Signaling Solutions,

Charlottesville, VA.] Bead-based antigen capture assays allow the analysis of up to 100 protein types per experiment. While commercial multiplex bead-based immunoassays are not well suited for broad-based marker discovery, they do provide well-standardized quantitative assays for 10-100 proteins per sample, they include many proteins known to be associated with infection and immunity, and the assays can be completed within a few hours from sample collection.

The primary objective of these studies is to determine if immunoassays can be used to differentiate sick animals from healthy animals prior to overt signs of disease. Other key questions include providing an initial assessment of 1) the number and types of markers responding to disease; 2) the magnitude of concentration changes in these markers; 3) the time course of marker changes before, during, and after overt disease; and 4) the significance of these disease-induced changes compared with natural variation among uninfected mice.

Materials and Methods

Virus and mouse stocks

A TK- strain of cowpox virus (CPV) was used for these studies. CPV (TK- Brighton strain), Stock# 010605 was used for Experiment #1, with a target inoculum of 10^7 pfu/mouse. The same viral strain was used for Experiments #2, but the target inoculum was reduced to 10^6 pfu/mouse because of the large number of animals used in the second study. The time course of disease appeared identical for the two experiments although the symptoms appeared somewhat less severe in the second study. Female BALB/c mice (Harlan, Specific Pathogen Free) were used for all experiments with the age at challenge of 52-74 days. Plate counts were used to determine CPV concentrations in both the stock solutions, and in the lungs of three animals per experiment.

Inoculation and sample collection protocols

All experimental work with the mice was performed in a biosafety level 3 (BSL-3) animal facility at the University of New Mexico [11,7]. Mice were infected with CPV by surgical intratracheal instillation. The infected mouse group received 50 ul of media (PBS 2.5% BGS diluted 1:2 in Tris buffer) containing CPV. The sham group received the same intratracheal surgical procedure with 50ul of media only. The naïve control group received no surgery or CPV.

Animals were sacrificed at each experimental time point to obtain serum samples. The animals were euthanized via Avertin overdose. Blood was collected from clipped vena cava and placed in eppendorf tube. The blood was allowed to coagulate for at least 30 minutes at 4°C. The whole blood was then centrifuged at 4500 RPM for 10 min in microfuge, and the serum from each mouse was placed in individual eppendorf tubes. The samples were stored at -80°C, and shipped overnight on dry ice to LLNL at end of

study. Serum volumes varied from 700 to 200 microliters, with lower recoveries at the peak of the disease.

Design of the two experiments

The first experiment was designed to study the time course of disease from 3 hours to 10 days post inoculation. Seven CPV-infected mice and 3 sham-infected mice were sacrificed at each of the following time points post-inoculation: hour 3, and days 1, 3, 6, 8, and 10. Ten untreated naive mice were sacrificed at the same time as the day 8 mice. Finally 3 additional mice that received CPV inoculation were sacrificed immediately after CPV instillation to measure the viable CPV concentration in the inoculum.

The second experiment was designed to have larger numbers of animals in the time period before clinical signs (days 2 through 6). Experiment #2 was split into two identical groups because of the difficulty in processing all the animals at the same time. The protocol for each group was the following: Five CPV-infected mice and five sham-infected mice were sacrificed at each of the following time points post-inoculation: days 2, 3, 4, 5, 6 and 8. Five untreated naive mice were sacrificed at day 2, and five additional naive mice were sacrificed at day 8. Three mice were again used to assess the concentration of the inoculum. Thus, the complete experiment #2 had a total of 10 infected mice and 10 sham mice at each time point, and 20 naive mice. Experiment #2 also included a Survival Group that was maintained for the full 10 days of the study to monitor the time course of the infection. Infected (5), Sham (5) and Naïve (5) animals were ear marked and weights were measured for each of the animals were performed daily from day 0 through day 10 of the study. Visual observations of changes in appearance and activity were also recorded. Behavior changes were recorded using the following numerical categories: 0 = No signs of sickness; -1 = Signs of decreased activity; -2 = Mice appear very sick, ruffled hair, decreased activity; -3 = Non-responsive, moribund. No animals in the study reached the –3 moribund category. While more difficult to quantify, these behavioral changes generally correlated with body weight changes due to the infection.

Multiplex immunoassays of mouse serum proteins

Mouse serum samples were transported frozen on dry ice to Rules Based Medicine (RBM) for multiplex immunoassay analysis (Rules Based Medicine, Inc., Austin, TX). RBM utilizes a multi-analyte panel to quantify the concentrations of about 60 host antigens in mouse serum samples. A total of 50 microliters of serum were shipped for each analysis, with all serum samples coded for blind analysis. While some aspects of this multiplex analysis are proprietary, the following briefly describes the approach. Sample is incubated with a mixture of fluorescently labeled microsphere types, with each bead type conjugated to a different capture antibody. A mixture of biotinylated secondary antibodies is then added to label bead-captured antigen. Finally streptavidin-phycoerythrin is added to fluorescently label the captured antigen. . Flow cytometric analysis with a Luminex 100 flow analyzer is used to quantify fluorescence signals for ach antigen in the analysis. Purified antigen standards are included is some samples to

develop standard curves for relating bead-based fluorescence to antigen concentration. Sample processing typically requires 1 to 3 hours, and flow cytometric analysis takes about 1 minute per sample [16,17,24]. Additional details of the RBM analysis can be found at www.rulesbasedmedicine.com .

Assay results are reported in units of protein concentration. A "Lowest Detectable Dose" (LDD) is also reported for each antigen. The LDD is the antigen concentration that produces a fluorescence signal that is 3 standard deviations above the fluorescence of negative control beads. Bead fluorescence valued below the LDD can be measured, but these results are generally noisier and less reliable than signals above the LDD. See Table 1 for examples of signals above and below the LDD. Two antigens IL-6 and KC produced no detectable fluorescence in serum samples from normal mice. These antigens were, however, well above the LDD for some treated animals. For these antigens the lowest observed concentrations were used as an upper limit to allow calculations of fold changes between treated and control. The minimum observed concentration for these antigens were 8.6 pg/ml for IL-6, and 0.04 ng/ml for KC.

Results

Serum protein levels in naïve mice

The serum analyte panel used for these studies continues to be developed with the addition of new assays and the improvement of existing assays. Of the 60 proteins in the current panel, 4 assays were modified and 3 assays were added since the beginning of these studies. Table 1 lists the 53 proteins that were included in both experiments in this study, and experimental data from 8 naive mice and 8 replicate serum samples from Experiment #1. The markers are presented in two groups. The top group consists of markers with concentrations above the least detectable dose (LDD) in naïve animals. The LDD is defined as the concentration producing an assay signal greater than three standard deviations above the value of negative control assays. The bottom group includes markers below the LDD in naïve mice. While analyte signals below the LDD can be measured, these measurements generally show reduced precision.

The data in Table 1 show the wide dynamic range of bead-based competitive immunoassays. Analyte concentrations in naïve animals vary over 7 orders of magnitude from 150 micrograms per ml. for Apo A1, to 13 picograms per ml. for EGF. Assay sensitivity varies dramatically among analytes depending on the quality of the antibody reagents. Apo A1 and EGF, for example show similar measurement precision among animals, and among replicate samples. Overall, precision values are typically between 10 and 20% for markers above the LDD, while precision values of 20 to 60% are seen for the lower markers. Interlukins (e.g. IL-2 to IL-17) are generally lower then the LDD, while chemokines (MCP, MDC, and MIP) are often well above the LDD in naive mice.

Clinical course of CPV infection in BALB/c mice

Past experience with this CPV TK- strain in BALB/c mice showed the following time course of disease after inoculation: signs of illness begin at day-6, peak disease at day-8, and the beginning of recovery at day-10. Plaque assays showed no viable virus in serum at any of these time points consistent with the infection being localized in lung tissue. While limited viral stocks resulted in two different inoculum concentrations being used for the two experiments, visual observation of the animals was consistent with the same time course of disease in both experiments. No deaths were observed for any of the CPV-infected animals consistent with previous observations that these doses are well below the lethal dose for BALB/c mice.

A cohort of animals was included in Experiment #2 to quantify the time of onset of disease in this experiment. Ten CPV-inoculated, and ten sham-inoculated were ear marked and weighed each day for 10 days post treatment. These animals were also monitored for behavioral signs of illness. The results in Figure 1 show the mean, and standard deviation weight for the 10 animals in each group and the summary scores for the signs of illness. While there appears to be a few animals with decreased activity at day-5, it is not until day-6 that a clear weight difference is seen between these two groups. The maximum weight loss occurred at day-8, with no further loss up to day-10.

Serum protein profiles from 3 hours to 10 days post-inoculation from Experiment #1

The first experiment to study CPV infection in mice was designed to assess serum protein responses from the first 3 hours post-inoculation through the initial signs of recovery 10 days later. This experiment included 3 sham-infected and 7 CPV-infected animals at each time point, and 8 naive animals that received no treatment. Serum samples from each of the animals were collected in separate tubes, and all tubes were coded before analysis. Figure 2 shows the results for one analyte, MIP-1b, measured in serum samples from each animal in the experiment. Results from the 8 naive mice are shown on the left (N), while points corresponding to the 3 sham and 7 infected animals are shown at each time-point post inoculation (H-3 to D-10). Standard curves for each marker were used to convert immunoassay measurements to serum concentrations. Serum concentrations of MIP-1b in pg/ml are shown on the Y-axis of Figure 2. These data show that the concentration of this marker increases on day-6, reaches a maximum value on day-8 at the peak of the disease, and then declined on day-10 as the mice begin to recover. While considerable scatter is seen in the data points at the peak of the disease, there appears to be a clear difference between sham and infected from days 6-10.

Plots of the results from each of the serum markers were used to identify candidate markers that showed clear differences between sham and infected mice, or clear differences among time-points. Many candidate markers showed subtle differences among experimental groups, but these markers were not selected if these differences were not consistently observed among all points in the groups. While some of these subtle

differences may have biological significance, our initial analysis off candidate markers focused on clear differences like the differences between sham and infected animals shown in Figure 2.

Candidate markers selected for further analysis are shown in Table 2. Most candidate marker profiles fit into two major groups. A total of 12 markers showed peak concentrations at day-8, with a clear difference between sham and infected animals. The results in Figure 2 and Figure 3a and 3b show examples of these profiles. While the detailed shape of each profile shows some variation, all cases show a rise on day-6, and a fall on day-10 with minimal response in the sham group. These markers appear to mirror the onset of disease and recovery in the infected animals. A second group of 3 markers showed peak concentrations at hour-3 in both sham and infected animals. The data for IL-6, shown in Figure 3c, illustrate this profile. The observation that both the sham and infected animals showed similar responses at hour-3 suggests that this response results from the inoculation process, not the CPV. This "sham effect" is only seen immediately after the inoculation and could result from insult to the animals from the surgical trachiotomy, anesthesia, or fluid instillation in the lungs. Note that while the peak responses for these 3 markers occur at hour-3, weaker responses are seen during active infection (see the IL-6 response at day-6 in figure 3c). Similarly some markers that peak at day-8, also show a weaker "sham effect" at hour-3 (see MCP-1 responses in figure 3a). This suggests that while most markers primarily respond to inoculation or active infection, some are affected by both conditions.

Two additional candidate markers showed unique profiles not seen in other markers. While all the markers described above showed increases in concentration with either disease or surgery, leptin was the only marker that appeared to decrease in concentration with CPV infection. Considerable scatter of data points was observed for this marker, but a clear decrease was seen at day-8. Results in Figure 3d show the complex response profile observed for haptoglobin. While no concentration changes were observed at hour-3, both the sham and infected animals showed similar elevations at day 1-3. Haptoglobin levels remained elevated from days 6-10 for CPV infected animals, but these levels decreased to pre-treatment values for sham animals during the same period. This suggests haptoglobin may be exhibiting a delayed sham effect, followed by a second elevation due to CPV infection.

In summary, Table 2 shows that 17 markers with distinctive profiles were observed out of the 53 markers analyzed (Table 1). The magnitude of these concentration changes varied from 2-fold to 30-fold, with most changes smaller than a factor of ten. The p-values in Table 2 suggest that most of these changes are significant, even though these values have not been corrected for multiple comparisons. The use of multiple animals per group with separate analyses of each animal facilitates identification of candidate markers using either visual identification or statistical comparisons. The dramatic sham effect at hour-3 demonstrates the importance of including sham controls for surgical intratracheal instillation, and shows that large changes in serum protein levels can be detected within 3 hours after treatment. The only consistent responses to CPV inoculation, compared to sham, were observed in the time period of visible symptoms of disease from days 6-10.

Except for the unique profile of haptoglobin, no consistent sham effects were observed after hour-3, and no consistent CPV infection effects were seen before day-6. The second experiment in this study was designed to include days 4-5, which were not done in the first experiment, and to include a larger number of animals to facilitate analysis of this period before the appearance of visible symptoms in the CPV-infected mice.

Detailed analysis of serum protein profiles on days 2 through 8 from Experiment #2

A second experiment was performed to focus on marker responses in the period from day-2 to day-8. Several adjustments in protocol were needed because of the large numbers of animals required for this study. Early time-points at hour-3 and day-1 were not included to allow staff to focus on the large numbers of surgical procedures on the first day of the experiment. Target CPV inoculum amounts were reduced 10-fold in the second experiment because of limited viral stocks. Visual observations of the mice and measurements of weight loss were consistent with the same time-course of disease in the two experiments, although the severity of the infections may have been smaller in the second experiment. Finally, the experiment was divided into two identical groups, with about a one-month gap between the two sets of surgeries and the two sets of immunoassay measurements. The data presented below consist of the results from both halves of the experiment. The complete experiment included 10 CPV-infected and 10 sham mice at each time-point (days 2, 3, 4, 5, 6, and 8), and a total of 20 naive mice. The complete experiment also included 10 CPV- infected and 10 sham-infected mice that were monitored for body weight changes for 10 days after surgery.

Individual measurements from all 140 animals for one marker, MCP-1, are shown in figure 4. The compact clusters observed for most groups demonstrate the reproducibility of both the animal experimentation and immunoassays over a period of about 2 months. Particularly notable are the points corresponding to naive animals, as these animals were sacrificed at 4 different time-points (day-2 and day-8 from each experiment). Visual observation indicate minimal differences between CPV and sham at days 2-3, but clear differences at days 6-8 consistent with the results from experiment #1. These results, however, show clear differences between CPV and sham for days 4-5. This indicates that elevations in MCP-1 level can be clearly observed for 2 days prior to the onset of disease-associated body weight changes at day-6.

Visual inspection was again used to identify candidate markers that differ between CPV- and sham-infected animals. The results of this analysis are presented in Table 3 and Figure 5. These candidates include several markers that show elevated concentrations in the early period from day 4 and 5, as well as in the period of active infection. These markers appear to fall into 4 different categories based on their expression profiles, suggesting variations in expression profiles during the course of infection. The first category contains markers that show the largest differences between CPV and sham on days 4-5, with some elevations seen at later times (Fig. 5a). The second category showed largest responses on days 5-6 (Fig. 5b). The third category showed elevated levels for the whole period from day 4 to 8, although the detailed profiles varied some among markers

(Fig. 4, Fig. 5d). The final category is markers showing their peak response at day 8 (Fig. 5c). In this experiment, 2 markers were identified that showed a decrease in expression with infection. These include leptin (as seen in Expt.1), and CD40 ligand (Fig. 5e), a new marker assay that was not available for the first experiment. Finally, haptoglobin again showed a biphasic response with elevations in CPV and sham at early times, followed by an elevation for CPV only at later times.

Diagnostic potential of serum proteins in CPV infected BALB/c mice

Overall, the identity of markers altered, and the time-course of response of these markers, was very consistent between the two experiments. No consistent changes were observed in the early time period from day 1-3 from both experiments. Most markers with altered expression on day 8 from the first experiment show alterations in the same direction over multiple days in the second experiment. The 3 markers that differ between the two experiments include MIP-2 and IFN-g that showed peak responses at hour-3 and day-4 respectively, and the new assay for CD40 ligand. Finally, although the magnitude of concentration changes reported in Table 3 are frequently only 2- 5-fold, the p-values associated with these numbers confirm the ability of immunoassays to reliably quantify subtle changes in concentration.

The results from experiment #2 were also used to determine the potential of individual markers to identify animals at different stages of infection. First markers were evaluated for their ability to differentiate animals from days 4-8 after infection ("disease" n=40), from all other animals including naive, sham, and infected before day 8 ("non-disease" n=100). Table 4a shows results for 5 different markers. Concentration thresholds were manually selected to minimize both false positives and false negatives. MCP-1 showed the best discrimination, correctly identifying all 40 "disease" animals, and having only 1 false positive out of 100 "non-disease" animals. Thus, the disease effect (4- to 14-fold for MCP-1) is clearly larger than variations in the "non-disease group due to variations among experiments, variation among animals, sham effects, and infection effects before day 4. While the discrimination between "disease" and "non-disease" is somewhat reduced for the other 4 markers in Table 4a, specificity and sensitivity values are generally greater than 90 percent for all 5 markers.

The varying time-course of responses among markers shown in Table 3 suggests that multiple markers may also be used to distinguish different stages of disease. One example of this type of analysis is shown in Table 4b. MCP-1 was used to discriminate "disease" from "non-disease". An early marker of infection, IP-10, was then used to discriminate "disease" day 4-5 from "disease" day 6-8. While discrimination between these two time periods is more difficult, the results in Table 4b show that 75-80% of the animals are correctly classified. Even this limited success is remarkable, since animals show no signs of disease on day 4, and minimal signs of reduced activity on day 5. The one false positive animal from Table 4a was clustered with the day 4-5 animals using this analysis. Additional approaches for multivariate analyses using these data are in progress to assess their utility in characterizing different stages of disease progression.

Discussion

There is growing interest in the potential use of biochemical components in blood to provide indicators of disease [12,28,31]. While the current study is focused on infectious disease, similar approaches are being evaluated for detection and characterization of other physiological states including cancer [33], chronic diseases [32], ionizing radiation exposure [9] , and physical fitness [22,23]. A key hypothesis underlying all these studies is that biochemical signatures exist that differentiate each of these diseases from normal background variability. Given the complexity and interrelationships among physiological pathways in different diseases, precise measurements of large numbers of protein and/or RNA species may be required to produce a robust biochemical signature.

The present mouse study was designed to provide initial quantitative data on the systemic effects of an infectious disease in whole animals. The results in Figure 1 demonstrate that this disease model has a clear incubation period of 5-6 days, followed by signs of serious disease at about day 8, followed by initial signs of recovery by day 10. This model also allows assessment of systemic host response in peripheral blood from an infection that is localized in the lung. Finally the small size of mice makes it feasible to include many replicate animals at each time point strengthening statistical analysis of disease effects.

Bead-based multiplex immunoassay of blood proteins was selected to provide biochemical signatures in the present pilot study of CPV infection in mice. An important advantage of this approach is that commercial kits and analysis services are available for panels of 10-100 different antigens per analysis. The markers listed Table 1 show the wide variety of mouse serum markers in the present study including protein markers that have been associated with inflammation and disease (e.g. chemokines and cytokines), as well as other markers for the physiological status of the animal (e.g. clotting factors, lipid metabolism, acute phase reactions). The current marker set also allows initial assessment of the effects of technical factors, including limits of detection, variability among samples and animals, and signal-to-noise factors, on the interpretation of multiplex immunoassay studies. The coefficient of variation values among replicate experiments, and among animals are typically less the 30% for most markers (Table 1), confirming the high precision that can be obtained with this technology.

While a number of studies have reported changes of cytokines and chemokines in mouse models of viral infection, most of these studies performed analyses only on the tissues or fluids at the site of the infection. The first experiment in this study was designed to determine the nature, and timing, of protein responses in peripheral blood during the period from 3 hours to 10 days post inoculation with CPV. This period covers responses from immediately after surgical inoculation till the first signs of recovery from the infection. Two major patterns of protein responses were observed in this study as shown

in Table 2. Many serum markers showed clear elevations in concentration during the period of active disease (days 6-10). A second group of markers showed peak responses at 3 hours post inoculation.

Twelve markers showed clear elevations during active disease, peaking at day 8 when the mice appeared most ill. The magnitude of these concentration changes varied from 2-fold to 20-fold. Most of these markers are chemokines, and this cascade of chemokine production during the peak of disease is similar to responses seen with other viruses. Intranasal infection of mice with gammaherpes virus 68 resulted in elevated expression of 7 chemokines in lung tissue with peak expression at the time of peak viral load [3]. This study reported elevated chemokine expression up to 29 days post infection in contrast to our CPV results. Intranasal infection with vaccinia virus also produced a wave of cytokines and chemokines in bronchoalveolar lavage (BAL) fluids peaking at 10 days post infection. Significant elevations were seen for MCP-1, MIP-1alpha, eotaxin, IFN-gamma, and TNF-alpha [6]. Similar results have also been seen in the lungs of mice infected with respiratory syncytial virus [4]. Overall, it appears that cytokines and chemokines that are elevated in the lung are also clearly elevated in peripheral blood. We did not observe a few markers in blood, including TNF-alpha and MIP-1 alpha, but in both cases these proteins were below the limit of detection of our assay. Finally, while the quantitative profile of these markers may differ among viruses, none of these markers appears to be specific for one virus.

A second group of markers from experiment 1 showed peak responses 3 hours after inoculation, with similar responses seen for both sham and viral infected animals. This strongly suggests that this response results from the surgical intratracheal instillation protocol. This inoculation method allows precise control of exposure dose, but it clearly could induce injury responses in the animals. The pattern of markers showing this effect is very different from the disease associated markers seen at day 8. Recent reports have shown that lung damage can result in an influx of polymorphonuclear leukocytes in response to increased concentrations of IL-6, MIP-2, and KC [29,30]. This provides a possible mechanism for the surgery effect seen in this study, and supports the importance of including sham-treated animals in such studies.

Haptoglobin showed a unique and unexpected expression profile in both experiments 1 and 2 (Figs. 3 and 5). Haptoglobin concentrations increased for both infected and sham animals 1-2 days after infection. As the disease developed (days 6-10) a second concentration peak was observed for the infected animals, while the sham animals returned to near baseline levels. Haptoglobin has a strong binding affinity with hemoglobin. It is produced mainly in the liver, and is secreted into serum in response to inflammation, infection or trauma [1]. Thus, an increase in infected animals, but not shams, during active disease is not unexpected. Recent literature reports provide clear evidence that haptoglobin can be produced by some cell types in the lung, and that haptoglobin may play a key role in removing hemoglobin after lung injury [1,2]. Thus, the early elevation in haptoglobin in both infected and sham animals could result from surgical injury.

Experiment #2 was provided the first data on days 4 and 5, as these time points were not included in the first study. Additionally, this experiment was expanded to provide a balanced design of 10 infected and 10 sham animals at each time point. The objective of this experiment was to determine the time-course of marker responses from early in the incubation period through to active disease. The summary of results in Table 3 and Figure 5 show two major conclusions. First, while the important disease associated markers remained the same as in experiment 1, the second study showed clear evidence for elevations in several markers before the onset of any clinical signs of disease. Second, the study showed that the markers clustered into groups with different temporal profiles.

Two measures were used to try to define the first indications of disease in these animals. Reduction in body weight is an objective measure of illness that showed the first significant decreases 6 days after infection. Visual observations of behavior changes in the mice are more subjective, but they may be more sensitive to early effects. These observations showed "signs of decreased activity" in half of the mice on day 5, with the rest appearing normal until day 6. The observations reported in Table 3 show many markers are clearly elevated on days 4 and 5 after infection. Days 2 and 3, in contrast show no clear elevations consistent with the results from Experiment 1. It is striking that such a clear early response can be detected in peripheral blood rather that at the site of the infection. One study of neurological disease in mice induced by polytropic murine retroviruses showed initial elevations of cytokine and chemokine levels in the brain before major clinical symptoms consistent with early detection [10]. Studies with a series of virulent and attenuated strains provided evidence that cytokine/chemokine production was required for disease, suggesting that these markers may be involved in the pathology of the virus. It is unclear why these results, differ from the late appearance of markers in lung tissue described previously [3,4,6]. The high precision of bead-based immunoassays coupled with the use of multiple replicate mice may have led to increased power for early detection in serum seen in this study.

While these experiments do not address the question of discriminating among different disease types, the results of experiment 2 can be used to assess the reliability of this approach for early detection of CPV in mice. The results inn Table 4 show sensitivity and specificity values of 88 to 100% in identifying CPV infection as early as day 4 after infection. This shows that the "signal" of CPV infection from day 4-8 is well above the "noise" of experimental and biological variability in the model.

In conclusion, multiplex bead-based immunoassays are well suited for quantitative analysis of up to 100 protein types in the complex mixture of serum samples. They have sufficient accuracy to differentiate disease from healthy at the individual animal level. Additional assay types can be easily incorporated into the panel as new markers are discovered. Sample preparation and analysis can be completed within about 3 hours. The mouse CPV infection model has demonstrated that host responses are detectable in serum from a localized lung infection. These host responses can be reliably detected prior to overt signs of illness in the animals. Finally, the temporal pattern of response varies

among markers raising the possibility of using marker patterns to determine stages of infection

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48. Support for this work was provided by Laboratory Directed Research and Development funds.

References

1 Yang F, Ghio AJ, Herbert DC, et al. Pulmonary expression of the human haptoglobin gene. Am J Respir Cell Mol Biol 2000; 23: 277-282.

2 Yang F, Haile DJ, Berger FG, et al. Haptoglobin reduces lung injury associated with exposure to blood. Am J Physiol Lung Cell Mol Physiol 2003; 284: L402-L409.

3 Weinberg JB, Lutzke ML, Efstathiou S, Kunkel SL, Rochford R. Elevated chemokine responses are maintained in lungs after clearance of viral infection. J Virol 2002; 76: 10518-10523.

4 Miller AL, Bowlin TL, Lukacs NW. Respiratory syncytial virus-induced chemokine production: Linking viral replication to chemokine production in vitro and in vivo. J Infect Dis 2004; 189: 1419-1430.

5 Martinez MJ, Bray MP, Huggins JW. A mouse model of aerosol-transmitted orthopoxviral disease. Arch Path Lab Med 2000; 124: 362-377.

6 Reading PC, Smith GL. A kinetic analysis of immune mediators in the lungs of mice infected with vaccinia virus and comparison with intradermal infection. J Gen Virol 2003; 84:1973-1983.

7 Talaat AM, Lyons R, Howard ST, Johnston SA. The temporal expression profile of mycobacterium tuberculosis infection in mice. Proc Natl Acad Sci USA 2004; 101:4602-4607.

8 Rubins KH, Hensley LE, Jahrling PB, Whitney AR, Geisbert TW, Huggins JW, Owen A, Leduc JW, Brown PO, Relman DA.
Related Articles, Links

The host response to smallpox: analysis of the gene expression program in peripheral blood cells in a nonhuman primate model.
Proc Natl Acad Sci U S A. 2004 Oct 19;101(42):15190-5. Epub 2004 Oct 11.

PMID: 15477590 [PubMed - indexed for MEDLINE]

9 Coleman MA, Yin E, Peterson LE, Nelson D, Sorensen K, Tucker JD, Wyrobek AJ.
Related Articles, Links

Low-dose irradiation alters the transcript profiles of human lymphoblastoid cells including genes associated with cytogenetic radioadaptive response.
Radiat Res. 2005 Oct;164(4 Pt 1):369-82.
PMID: 16187739 [PubMed - indexed for MEDLINE]

10 Peterson KE, Robertson SJ, Portis JL, Chesebro B.
Related Articles, Links

Differences in cytokine and chemokine responses during neurological disease induced by polytropic murine retroviruses Map to separate regions of the viral envelope gene.
J Virol. 2001 Mar;75(6):2848-56.
PMID: 11222710 [PubMed - indexed for MEDLINE]

11 Lyons CR, Lovchik J, Hutt J, Lipscomb MF, Wang E, Heninger S, Berliba L, Garrison K.
Related Articles, Links

Murine model of pulmonary anthrax: kinetics of dissemination, histopathology, and mouse strain susceptibility.
Infect Immun. 2004 Aug;72(8):4801-9.
PMID: 15271942 [PubMed - indexed for MEDLINE]

12 Jahrling PB, Hensley LE, Martinez MJ, Leduc JW, Rubins KH, Relman DA, Huggins JW.
Related Articles, Links

Exploring the potential of variola virus infection of cynomolgus macaques as a model for human smallpox.
Proc Natl Acad Sci U S A. 2004 Oct 19;101(42):15196-200. Epub 2004 Oct 11.
PMID: 15477589 [PubMed - indexed for MEDLINE]

13 Shen Y, Tolic N, Zhao R, Pasa-Tolic L, Li L, Berger SJ, Harkewicz R, Anderson GA, Belov ME, Smith RD.
Related Articles, Links

High-throughput proteomics using high-efficiency multiple-capillary liquid chromatography with on-line high-performance ESI FTICR mass spectrometry.
Anal Chem. 2001 Jul 1;73(13):3011-21.
PMID: 11467548 [PubMed - indexed for MEDLINE]

14 Langlois RG, Trebes JE, Dalmasso EA, Ying Y, Davies RW, Curzi MP, Colston BW Jr, Turteltaub KW, Perkins J, Chromy BA, Choi MW, Murphy GA, Fitch JP, McCutchen-Maloney SL.
Related Articles, Links

Serum protein profile alterations in hemodialysis patients.
Am J Nephrol. 2004 Mar-Apr;24(2):268-74. Epub 2004 Mar 19.
PMID: 15031630 [PubMed - indexed for MEDLINE]

15 Fodor IK, Nelson DO, Alegria-Hartman M, Robbins K, Langlois RG, Turteltaub KW, Corzett TH, McCutchen-Maloney SL.
Related Articles, Links

Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyder.
Bioinformatics. 2005 Oct 1;21(19):3733-40. Epub 2005 Aug 9.
PMID: 16091413 [PubMed - in process]

16 McBride MT, Masquelier D, Hindson BJ, Makarewicz AJ, Brown S, Burris K, Metz T, Langlois RG, Tsang KW, Bryan R, Anderson DA, Venkateswaran KS, Milanovich FP, Colston BW Jr.
Related Articles, Links

Autonomous detection of aerosolized Bacillus anthracis and Yersinia pestis.
Anal Chem. 2003 Oct 15;75(20):5293-9.
PMID: 14710805 [PubMed - indexed for MEDLINE]

17 McBride MT, Gammon S, Pitesky M, O'Brien TW, Smith T, Aldrich J, Langlois RG, Colston B, Venkateswaran KS.
Related Articles, Links

Multiplexed liquid arrays for simultaneous detection of simulants of biological warfare agents.
Anal Chem. 2003 Apr 15;75(8):1924-30.
PMID: 12713052 [PubMed - indexed for MEDLINE]

18 Quenelle DC, Collins DJ, Wan WB, Beadle JR, Hostetler KY, Kern ER.
Related Articles, Links

Oral treatment of cowpox and vaccinia virus infections in mice with ether lipid esters of cidofovir.
Antimicrob Agents Chemother. 2004 Feb;48(2):404-12. Erratum in: Antimicrob Agents Chemother. 2004 May;48(5):1919.
PMID: 14742188 [PubMed - indexed for MEDLINE]

19 Quenelle DC, Collins DJ, Kern ER.
Related Articles, Links

Efficacy of multiple- or single-dose cidofovir against vaccinia and cowpox virus infections in mice.
Antimicrob Agents Chemother. 2003 Oct;47(10):3275-80.
PMID: 14506041 [PubMed - indexed for MEDLINE]

20 Chromy BA, Gonzales AD, Perkins J, Choi MW, Corzett MH, Chang BC, Corzett CH, McCutchen-Maloney SL.
Related Articles, Links

Proteomic analysis of human serum by two-dimensional differential gel electrophoresis after depletion of high-abundant proteins.
J Proteome Res. 2004 Nov-Dec;3(6):1120-7.
PMID: 15595720 [PubMed - indexed for MEDLINE]

21 Chromy BA, Choi MW, Murphy GA, Gonzales AD, Corzett CH, Chang BC, Fitch JP, McCutchen-Maloney SL.
Related Articles, Links

Proteomic Characterization of Yersinia pestis Virulence.
J Bacteriol. 2005 Dec;187(23):8172-80.
PMID: 16291690 [PubMed - in process]

22 Suzuki K, Nakaji S, Yamada M, Liu Q, Kurakake S, Okamura N, Kumae T, Umeda T, Sugawara K.
Related Articles, Links

Impact of a competitive marathon race on systemic cytokine and neutrophil responses.
Med Sci Sports Exerc. 2003 Feb;35(2):348-55.
PMID: 12569227 [PubMed - indexed for MEDLINE]

23 Pedersen BK, Steensberg A, Fischer C, Keller C, Ostrowski K, Schjerling P.
Related Articles, Links

Exercise and cytokines with particular focus on muscle-derived IL-6.
Exerc Immunol Rev. 2001;7:18-31. Review.
PMID: 11579746 [PubMed - indexed for MEDLINE]

24 Prabhakar U, Eirikis E, Davis HM.
Related Articles, Links

Simultaneous quantification of proinflammatory cytokines in human plasma using the LabMAP assay.
J Immunol Methods. 2002 Feb 1;260(1-2):207-18.
PMID: 11792390 [PubMed - indexed for MEDLINE]

25 de Jager W, te Velthuis H, Prakken BJ, Kuis W, Rijkers GT.
Related Articles, Links

Simultaneous detection of 15 human cytokines in a single sample of stimulated peripheral blood mononuclear cells.
Clin Diagn Lab Immunol. 2003 Jan;10(1):133-9.
PMID: 12522051 [PubMed - indexed for MEDLINE]

26 Pang S, Smith J, Onley D, Reeve J, Walker M, Foy C.
Related Articles, Links

A comparability study of the emerging protein array platforms with established ELISA procedures.
J Immunol Methods. 2005 Jul;302(1-2):1-12.
PMID: 15993890 [PubMed - indexed for MEDLINE]

27 Jiang B, Snipes-Magaldi L, Dennehy P, Keyserling H, Holman RC, Bresee J, Gentsch J, Glass RI.
Related Articles, Links

Cytokines as mediators for or effectors against rotavirus disease in children.
Clin Diagn Lab Immunol. 2003 Nov;10(6):995-1001.
PMID: 14607858 [PubMed - indexed for MEDLINE]

28 Boldrick JC, Alizadeh AA, Diehn M, Dudoit S, Liu CL, Belcher CE, Botstein D, Staudt LM, Brown PO, Relman DA.
Related Articles, Links

Stereotyped and specific gene expression programs in human innate immune responses to bacteria.
Proc Natl Acad Sci U S A. 2002 Jan 22;99(2):972-7.
PMID: 11805339 [PubMed - indexed for MEDLINE]

29 Lomas-Neira J, Chung CS, Perl M, Gregory S, Biffl W, Ayala A.
Related Articles, Links

Role of Alveolar Macrophage & Migrating Neutrophils in Hemorrhage Induced Priming for ALI Subsequent to Septic Challenge.
Am J Physiol Lung Cell Mol Physiol. 2005 Sep 9; [Epub ahead of print]
PMID: 16157517 [PubMed - as supplied by publisher]

30 Rijneveld AW, van den Dobbelsteen GP, Florquin S, Standiford TJ, Speelman P, van Alphen L, van der Poll T.
Related Articles, Links

Roles of interleukin-6 and macrophage inflammatory protein-2 in pneumolysin-induced lung inflammation in mice.
J Infect Dis. 2002 Jan 1;185(1):123-6. Epub 2001 Nov 30.
PMID: 11756992 [PubMed - indexed for MEDLINE]

31 Mahanty S, Gupta M, Paragas J, Bray M, Ahmed R, Rollin PE.
Related Articles, Links

Protection from lethal infection is determined by innate immune responses in a mouse model of Ebola virus infection.
Virology. 2003 Aug 1;312(2):415-24.
PMID: 12919746 [PubMed - indexed for MEDLINE]

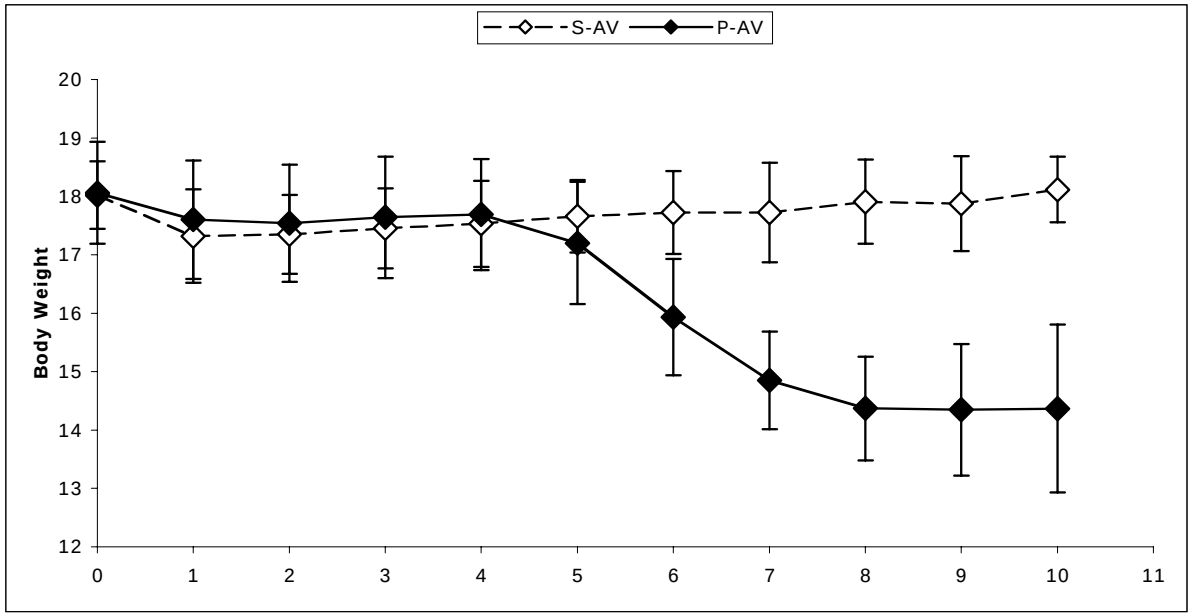
32 Tarakcioglu M, Erbagci AB, Usalan C, Deveci R, Kocabas R.
Related Articles, Links

Acute effect of hemodialysis on serum levels of the proinflammatory cytokines.
Mediators Inflamm. 2003 Feb;12(1):15-9.
PMID: 12745544 [PubMed - indexed for MEDLINE]

33 Poon TC, Yip TT, Chan AT, Yip C, Yip V, Mok TS, Lee CC, Leung TW, Ho SK, Johnson PJ.
Related Articles, Links

Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes.
Clin Chem. 2003 May;49(5):752-60.
PMID: 12709366 [PubMed - indexed for MEDLINE]

Figure 1


S-AV
P-AV
Body Weight
20
19
18
17
16
15
14
13
12
0
1
2
3
4
5
6
7
8
9
10
11

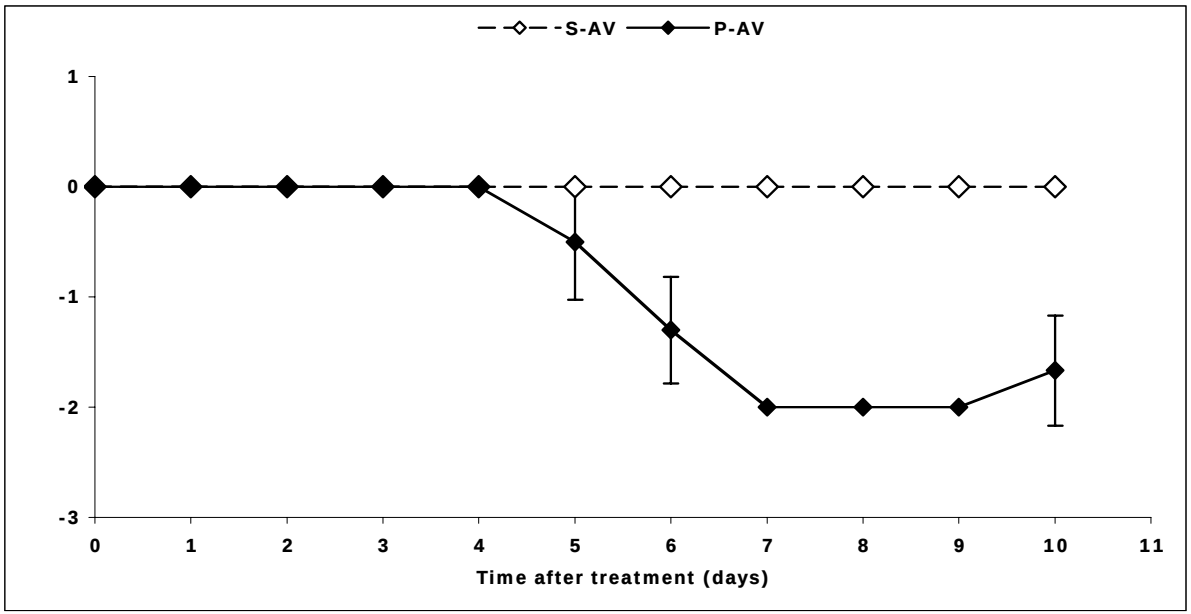

S-AV
P-AV
1
0
-1
-2
-3
0
1
2
3
4
5
6
7
8
9
10
11
Time after treatment (days)

Figure 2

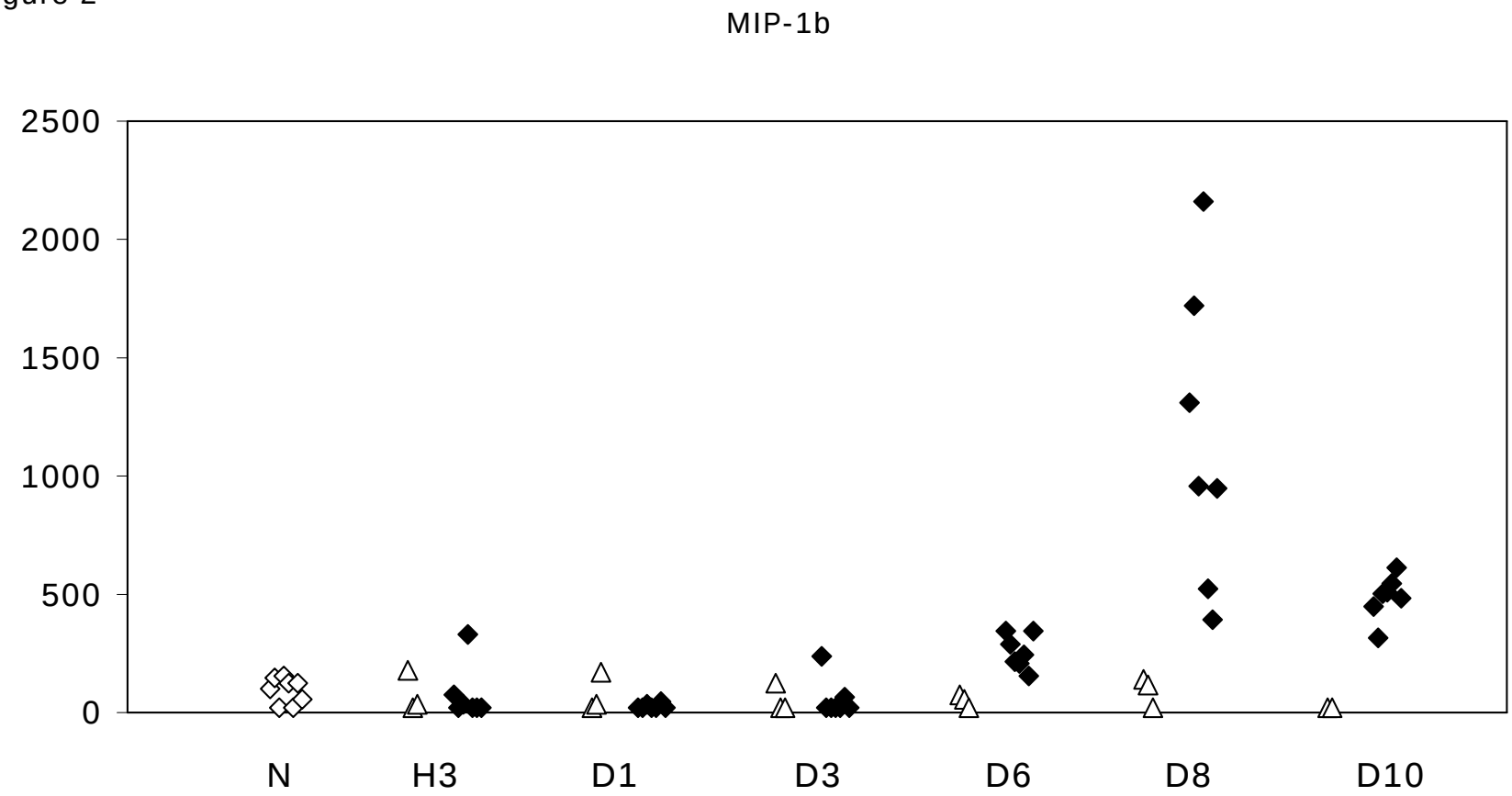


Figure 3

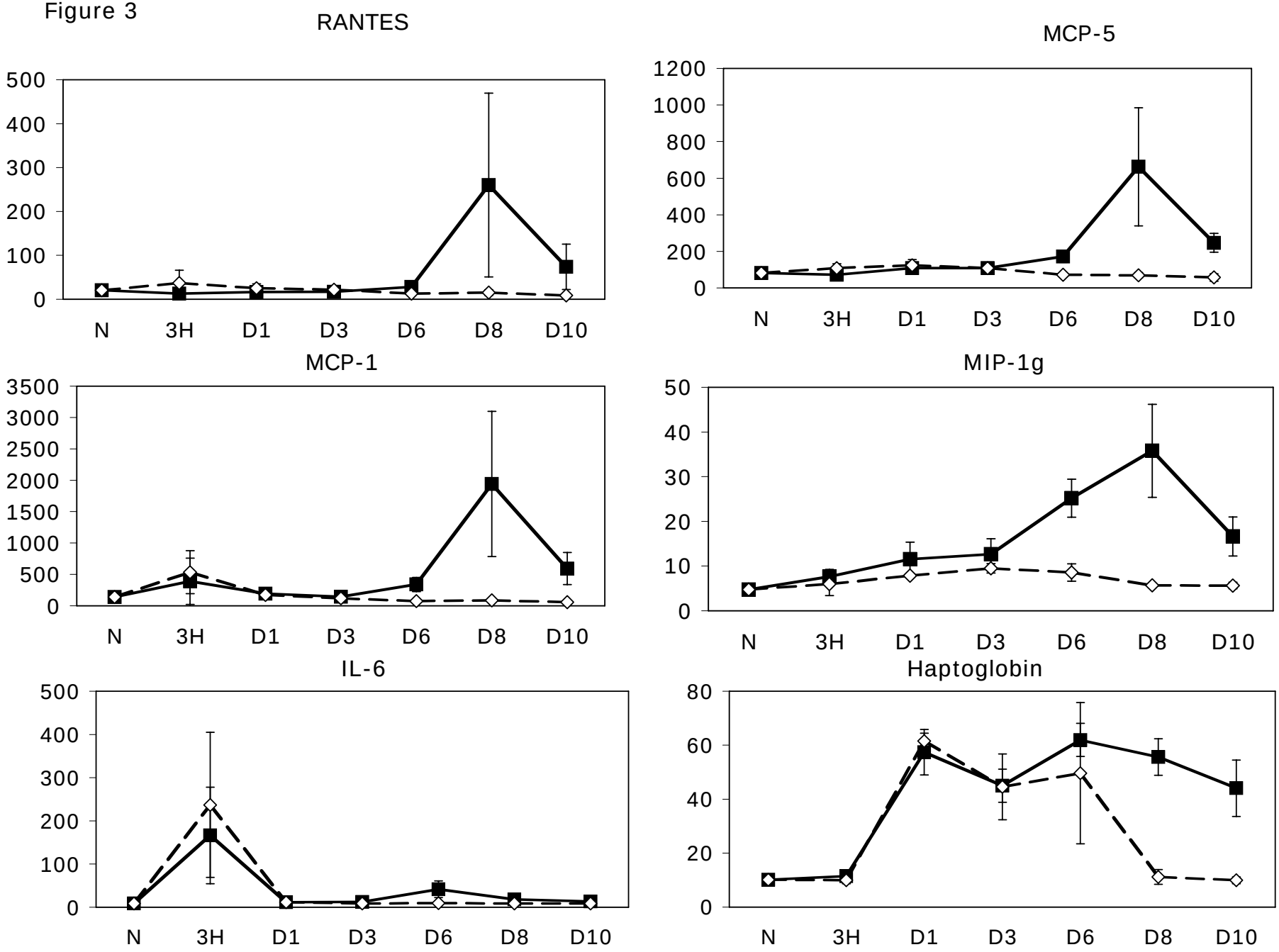

RANTES
500
400
300
200
100
0
N
3H
D1
D3
D6
D8
D10
MCP-5
1200
1000
800
600
400
200
0
N
3H
D1
D3
D6
D8
D10
MCP-1
3500
3000
2500
2000
1500
1000
500
0
N
3H
D1
D3
D6
D8
D10
MIP-1g
50
40
30
20
10
0
N
3H
D1
D3
D6
D8
D10
IL-6
500
400
300
200
100
0
N
3H
D1
D3
D6
D8
D10
Haptoglobin
80
60
40
20
0
N
3H
D1
D3
D6
D8
D10

Figure 4

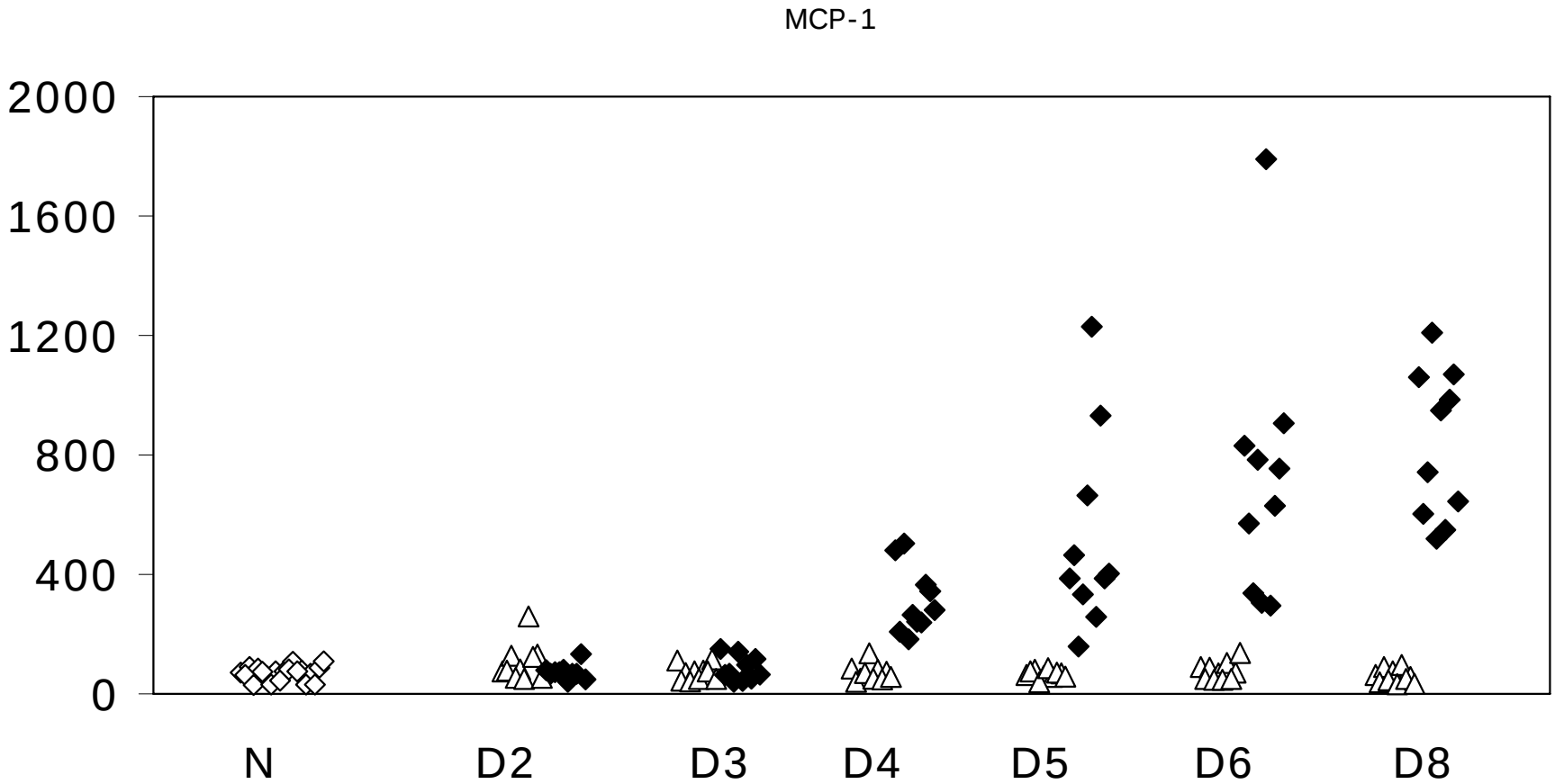


Figure 5

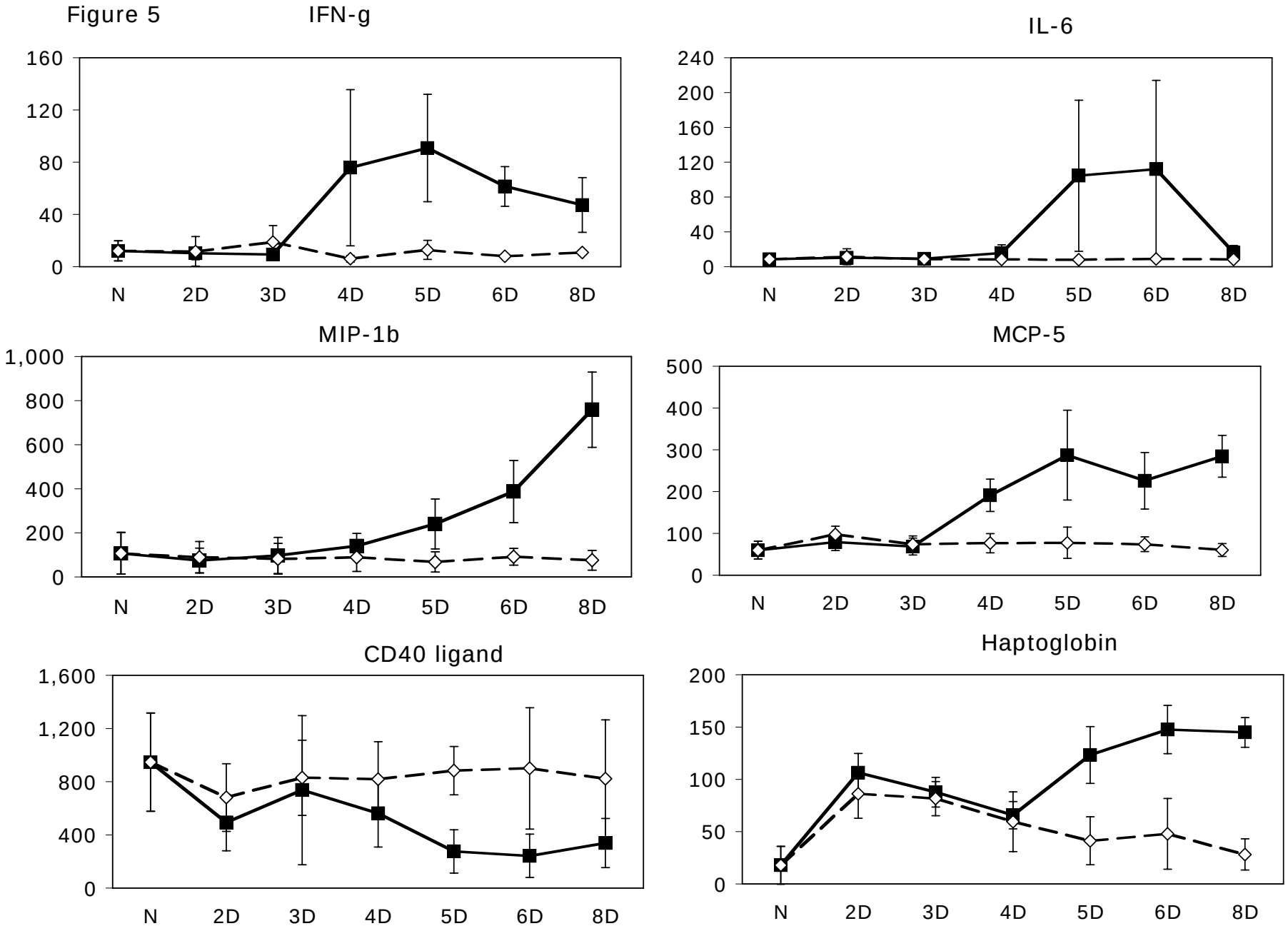

IFN-g
160
120
80
40
0
N
2D
3D
4D
5D
6D
8D
IL-6
240
200
160
120
80
40
0
N
2D
3D
4D
5D
6D
8D
MIP-1b
1,000
800
600
400
200
0
N
2D
3D
4D
5D
6D
8D
MCP-5
500
400
300
200
100
0
N
2D
3D
4D
5D
6D
8D
CD40 ligand
1,600
1,200
800
400
0
N
2D
3D
4D
5D
6D
8D
Haptoglobin
200
150
100
50
0
N
2D
3D
4D
5D
6D
8D

TABLE 1a Proteins above the Least Detectable Dose in naïve mice

Protein Symbol	units	LDD	naïve av/LDD	naïve av	naïve CV(%)	replicates CV(%)	Protein Name
Apo A1	ug/ml	10	14.47	150.50	7	9	Apolipoprotein A1
Eotaxin	pg/ml	12	52.34	641.13	23	5	Eotaxin
FGF-basic	ng/ml	0.58	1.16	0.67	35	23	Fibroblast Growth Factor-basic
GCP-2	ng/ml	0.025	557.14	13.65	15	9	Granulocyte Chemotactic Protein-2
Haptoglobin	ug/ml	0.64	15.77	10.09	10	7	Haptoglobin
IgA	ug/ml	1.9	33.21	62.76	15	9	Immunoglobulin A
IL-10	pg/ml	109	2.53	276.13	13	23	Interleukin-10
IL-18	ng/ml	0.67	1.06	0.72	29	22	Interleukin-18
IL-1alpha	pg/ml	45	6.85	307.88	18	15	Interleukin-1alpha
IP-10	pg/ml	40	5.87	236.31	109	18	Inducible Protein-10
Leptin	ng/ml	0.096	17.81	1.70	21	6	Leptin
LIF	pg/ml	44	1.14	49.88	12	13	Leukemia Inhibitory Factor
MCP-1	pg/ml	17	8.21	139.00	21	10	Monocyte Chemoattractant Protein-1
MCP-3	pg/ml	31	8.46	265.25	18	11	Monocyte Chemoattractant Protein-3
MCP-5	pg/ml	46	1.75	81.23	17	11	Monocyte Chemoattractant Protein-5
M-CSF	ng/ml	0.018	205.50	3.69	13	5	Macrophage-Colony Stimulating Fact
MDC	pg/ml	22	14.91	325.88	12	8	Macrophage-Derived Chemokine
MIP-1beta	pg/ml	78	1.34	103.81	47	47	Macrophage Inflammatory Protein-1b
MIP-1gamma	ng/ml	0.074	64.81	4.77	17	11	Macrophage Inflammatory Protein-1g
MIP-2	pg/ml	7.2	6.24	44.74	28	23	Macrophage Inflammatory Protein-2
Myoglobin	ng/ml	24	32.19	772.50	83	13	Myoglobin
SGOT	ug/ml	1.9	12.20	22.69	16	4	Serum Glutamic-Oxaloacetic Transan
TIMP-1	ng/ml	0.18	10.27	1.85	21	15	Tissue Inhibitor of Metalloproteinase
Tissue Factor	ng/ml	0.52	7.60	3.93	9	7	Tissue Factor
TPO	ng/ml	2.7	3.12	8.31	7	8	Thrombopoietin
VCAM-1	ng/mL	0.95	1166.80	1111.38	11	9	Vascular Cell Adhesion Molecule-1
VEGF	pg/ml	38	4.62	176.50	19	20	Vascular Endothelial Cell Growth Fac

TABLE 1b Proteins below the Least Detectable Dose in naïve mice

Protein Symbol	units	LDD	naïve av/LDD	naïve av	naïve CV(%)	replicates CV(%)	Protein Name
CRP	ug/mL	0.53	0.94	0.50	19	21	C Reactive Protein
EGF	pg/ml	39	0.34	13.46	21	11	Epidermal Growth Factor
Endothelin-1	pg/ml	67	0.16	10.88	34	-	Endothelin-1
Factor VII	ng/ml	0.96	0.97	0.93	14	21	Factor VII
FGF-9	ng/ml	0.99	0.28	0.27	32	-	Fibroblast Growth Factor-9
IFN-gamma	pg/ml	68	0.61	41.15	40	61	Interferon-gamma
IL-11	pg/ml	87	0.46	39.77	51	43	Interleukin-11
IL-12p70	ng/ml	0.57		nd			Interleukin-12p70
IL-17	ng/ml	0.15	0.15	0.02	43	52	Interleukin-17
IL-1beta	ng/ml	0.45	0.57	0.25	18	18	Interleukin-1beta
IL-2	pg/ml	67	0.24	16.05	47	82	Interleukin-2
IL-3	pg/ml	21	0.69	14.64	140	27	Interleukin-3
IL-4	pg/ml	74	0.41	30.05	40	46	Interleukin-4
IL-5	ng/ml	0.19	0.42	0.08	20	29	Interleukin-5
IL-6	pg/ml	14		nd			Interleukin-6
IL-7	ng/ml	0.31	0.23	0.07	34	30	Interleukin-7
Insulin	ulU/ml	2.0	0.64	1.26	57	26	Insulin
KC/GROalpha	ng/ml	0.17		nd			Melanoma Growth Stimulatory Activity Protein
Lymphotactin	pg/ml	85	0.93	79.53	22	23	Lymphotactin
MIP-1alpha	ng/ml	0.23	0.54	0.12	24	18	Macrophage Inflammatory Protein-1alpha
MIP-3beta	ng/ml	0.47	0.34	0.16	35	28	Macrophage Inflammatory Protein-3beta
OSM	ng/ml	0.13	0.37	0.05	50	46	Oncostatin M
RANTES	pg/ml	48	0.43	20.48	41	15	Regulation Upon Activation, Normal T-Cell Expresse
SCF	pg/ml	75	0.83	62.33	27	21	Stem Cell Factor
TNF-alpha	ng/ml	0.14	0.22	0.03	23	27	Tumor Necrosis Factor-alpha
vWF	ng/ml	99	0.17	17.13	62	-	von Willebrand Factor

TABLE 2- Proteins with altered concentrations in Experiment #1

Protein Label	Type*	Feature	3H/N	3H vs. N p-value	P8/S8	P8 vs. S8 p-value
IL-6	cytokine	3-Hours	21.8	0.001		
KC/GROalpha	CXC	3-Hours	36.1	0.011		
MIP-2	CXC	3-Hours	5.1	0.018		
IP-10	CXC	Day-8			6.6	0.014
TIMP-1		Day-8			4.1	>0.001
MCP-1	CC	Day-8			22.2	0.005
MCP-5	CC	Day-8			9.6	0.003
MCP-3	CC	Day-8			6.7	0.001
MIP-1gamma	CC	Day-8			6.3	>0.001
Lymphotactin	C	Day-8			3.2	0.004
IL-18	cytokine	Day-8			1.9	>0.001
Eotaxin	CC	Day-8			8.0	0.008
MIP-1beta	CC	Day-8			12.4	0.004
RANTES	CC	Day-8			14.2	0.022
IL-11	cytokine	Day-8			6.7	0.014
Leptin	wound	Negative			0.4	0.002
Haptoglobin	acute phase	Unique			5.0	>0.001

* chemokines listed by family (CXC, CC, C)

Table 3- Proteins with altered concentrations in experiment #2

Marker	Profile	Concentration ratio of infected to sham animals				
		day 2	day 3	day 4	day 5	day 6
IP-10	pk 4-5	0.7	1.1	*** 5.7	*** 3.7	*** 2.2
IFN-gamma	pk 4-5	0.9	0.5	** 12.5	*** 7.1	*** 7.7
TIMP-1	pk 4-5	0.9	1.0	*** 2.5	*** 2.4	* 1.4
Lymphotactin	pk 4-5	* 0.6	0.8	** 1.6	* 1.7	0.8
IL-6	pk 5-6	0.9	1.1	* 1.9	** 12.9	* 12.6
KC/GROalpha	pk 5-6	1.0	1.0	1.0	** 3.3	** 3.1
MCP-1	up 4-8	0.7	1.2	*** 4.5	*** 8.0	*** 9.6
MCP-5	up 4-8	0.8	0.9	*** 2.5	*** 3.7	*** 3.1
MCP-3	up 4-8	0.8	1.0	*** 2.2	*** 2.0	*** 2.3
MIP-1gamma	up 4-8	* 1.3	1.2	*** 1.9	*** 2.9	*** 2.6
IL-18	up 4-8	* 1.2	1.1	** 1.3	** 1.4	* 1.2
MIP-1beta	pk 8	0.8	1.2	1.6	*** 3.5	*** 4.2
RANTES	pk 8	1.1	1.1	1.2	** 2.2	* 2.6
Eotaxin	pk 8	0.8	0.9	1.1	1.4	* 1.9
IL-11	pk 8	* 0.5	1.2	* 1.6	2.0	1.9
CD40 Ligand	neg	0.7	0.9	* 0.7	*** 0.3	*** 0.3
Leptin	neg	*** 0.5	0.8	0.8	1.3	0.8
Haptoglobin	unique	* 1.2	1.1	1.1	*** 3.0	*** 3.1

Notes- p-values comparing infected vs. Sham

* ≤ 0.05, ** ≤ 0.01, *** ≤ 0.002

Table 4a- Protein markers characteristic of infection days 4-8

Protein	Threshold	TN=100 False Positives	TP=40 False Negatives	Sensitivity	Spe
MCP-1	158	1	0	100	
MCP-5	130	2	1	98	
MIP-1gamma	21	4	2	95	
IP-10	120	6	5	88	
IFN-gamma	21	4	2	95	

Table 4b- Protein markers differentiating early vs. late period of infection

Threshold MCP-1	Threshold IP-10	TP=20 Day 4-5	TP=20 Day 6-8	Sensitivity	Spe
≥158	≥201	16	5	80	
≥158	<201	4	15		