Quantitative linkage of physiology and gene expression through empirical model construction: an investigation of diabetes

J. Misra[1], I. Alevizos[1], J. Bullen[2], S. Blueher[2], C. Mantzoros[2], G. Stephanopoulos[1,*]

This work was supported by the Singapore-MIT Alliance.

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Division of Endocrinology, Beth Israel Deaconess Medical Center, Boston, MA 02215
* Corresponding author. Email: gregstep@mit.edu

Abstract— A methodology for the construction of predictive empirical models of physiological characteristics from microarray data is presented. The method, applied here to the study of the development of diabetes and insulin resistance, can be further expanded to other cases and to also include a variety of other data, such as protein expression, or metabolic flux data. The importance of several of the genes identified by the modeling methodology can be verified by comparison with results from prior literature. This implies potentially significant roles in diabetes for several of the uncharacterized genes discovered during the modeling procedure.

***Index Terms*—Diabetes, microarrays, partial least squares, systems biology**

I. INTRODUCTION

Gene expression measurements by DNA microarrays are often conducted either with a view to identify similarly expressed genes, or genes that are differentially regulated between two or more conditions under examination. Besides clustering [1] and classification [2], another important problem that arises when quantitative transcriptional data are available is the construction of models that have the ability to predict physiological variables from the values of gene expression as measured on microarrays. The importance of such models lies in the field of quantitative diagnostics, and toxicology models. The problem has received to date insufficient attention, primarily due to lack of a quantifiable phenotype, and the availability of a sufficient number of samples for reliable model construction. In this study, we designed an experiment that would provide a framework for the construction of these predictive models, not just in an ideal *in vitro* environment, but in a more complex *in vivo* situation.

Cohort experiments were conducted to track the development of diabetes in C57bl6 mice maintained on a high-fat diet for 12 weeks. Diabetes has the advantage of a quantifiable phenotype as measured through glucose and insulin levels in the body, as well as body fat %. Each week, mice were sacrificed, and the liver harvested for microarray analysis. Further, the serum was obtained to measure systemic glucose and insulin levels. Prior to sacrificing, DEXA measurements were conducted to obtain abdominal fat %. Based on the transcriptional profiles, empirical regression models were constructed between gene expression measurements and the physiological variables, and the predictive power of these models was evaluated. The modeling effort also identified a list of genes, several of which had previously been implicated in the development of diabetes, and several novel genes that bear potential for further investigation.

II. METHODS

A. Experimental

RNA isolation: The liver tissues were homogenized (1ml/50mg tissue) in RNA STAT-60 (Tel-Test, Friendswood, TX) with a Tissue-Tearor™ (Biospec Products, Bartlesville, OK). Following homogenization, the homogenate was stored for 5 minutes at room temperature to permit the complete dissociation of nucleoprotein complexes. 0.2ml of chloroform per 1 ml of RNA STAT-60 were then added and the mix was vigorously shaken for 15 seconds and centrifuged at 12,000g for 15 minutes at 4°C. After centrifugation the aqueous phase was transferred to a fresh tube and mixed with 0.5ml of isopropanol per 1ml of RNA STAT-60 used for the initial homogenization. The samples were stored at room temperature for 10 minutes and centrifuged at 12,000g for 10 minutes at 4°C. The supernatant was then removed and the pellet was washed with 1ml of 75% ethanol, dried and resuspended in Rnase free water. The RNA was further purified following the Rneasy Mini kit (Qiagen, Valencia, CA) protocol for RNA cleanup.

Microarrays:

i. Oligonucleotide library and Printing: The Operon Qiagen Mouse Genome Oligo Set (Operon Qiagen, Alameda, CA) Version 2 was used for the creation of the DNA microarrays. The set contains 16,463 *M. musculus* genes and 24 controls. The set was resuspended in 30µl of RNase and Dnase-free 3x SSC for a final concentration of

20µmol. The set was printed on Corning GAPS II coated barcoded slides (Corning, Corning, NY). Printing quality was assessed by SYBR II staining (Molecular Probes, Eugene, OR).

ii. Microarray validation: The microarrays were extensively validated for intra-array and inter-array variability. A mean coefficient of variation of 20% was observed based on repeat experiments. This implies that ratios greater than 1.4 and less than 0.6 can be considered significant at a 95% confidence limit. In addition, experiments were conducted to ensure that differential expression could be detected. Total RNA from healthy skeletal muscle and testes tissues were hybridized against each other, while on a control array skeletal muscle RNA was hybridized against itself. Based on our cut-offs established above, about 35% of the genes were found to be differentially expressed in the test array, as opposed to only 6% in the control array. Of the genes differentially expressed, several were specific to skeletal muscle, such as the skeletal muscle myosin, troponins, and actins.

iii. Control RNA: Control RNA for all the hybridizations was derived by pooling RNA from 20 mice from the following tissues: hypothalamus, liver, skeletal muscle, brown fat, white fat, kidney, adrenal gland, testis, ovary, heart and lung.

iv. Labeling and Hybridization: 10µg of RNA were used for both the control and the samples. The labeled cDNA synthesis took place as follows: 2 µL oligo-dT_{18-20} primer (Invitrogen, Carlsbad, CA) were added to the sample and the mix was heated to 70°C for 10 minutes, followed by 2 minute incubation on ice. Subsequently, 2.0 µL of 10X Cy3 dCTP (PerkinElmer, Boston, MA) for the control and 10X Cy5 dCTP (PerkinElmer) dCTP were added, followed by 2 µL 10X dNTPs (Invitrogen), 2 µL 100 mM DTT (Invitrogen), 4 µL 5X First Strand Buffer (Invitrogen), 2 µL Superscript II (Invitrogen). The final mix was then incubated at 42°C for 2 hours. After the end of the reverse transcription, 1.5 µL of 1 N NaOH were added in each sample and a further incubation took place at 65 °C for 10 minutes. Then, 1.5 µL of 1 N HCL were added to each sample to neutralize NaOH. The Cy3 and the Cy5 samples were combined and purified from the unincorporated dyes, nucleotides and enzymes with the Qiagen QIAquick nucleotide removal kit. The samples were then concentrated and resuspended in 20 µL of warm GlassHyb™ hybridization buffer (Clontech, Franklin Lakes, NJ) and applied on the microarray slides. A coverslip was placed on top of the slides with care not to create bubbles under the coverslip. The slides were subsequently sealed in Corning Hybridization Chambers (Corning) and left overnight in a covered water bath at 55°C for a total of 12 hours hybridization time. At the end of hybridization the coverslip was removed in a 1X SSC, 1% SDS solution and then washed for 5 minutes in the same solution, followed by a 5 minute wash in 0.2X SSC and a 5 minute wash in 0.1X SSC. The slides were then placed in a sterile Falcon tube and dried by spinning at 500xg for 3 minutes.

v. Scanning: The dried slides were scanned in the GenePix® 4000B microarray scanner (Axon, Union City, CA) and analyzed with the GenePix® Pro (Axon) acquisition and analysis software.

B. Partial least squares regression

Given a matrix **X** of independent variables, and a matrix **Y** of dependent variables, each with *s* samples, and *m*, and *n* variables respectively, partial least squares (PLS) can be used to develop a regression model between the two. PLS is particularly well suited for constructing regression models for microarray data, since the number of variables (genes) is much larger than the number of samples, and due to the large variation in microarray data. The approach consists of projecting the **X** and **Y** data matrices into a set of lower dimensions, or latent variables (LVs), and then constructing a regression model in this reduced space. The **X** matrix or block is linearly decomposed into a set of input scores, denoted by ***t***, and the **Y** matrix is similarly decomposed into a set of output scores, denoted by ***u***. These are known as the 'outer relations' in a PLS model. The regression between the ***t*** and ***u*** vectors is known as the 'inner relation'. Like the PCA, the decomposition is orthogonal, leading to successive ***t*** (or ***u***) vectors that are uncorrelated with their predecessors. Each ***t*** (or ***u***) vector is a linear combination of the input (or output) variables. In addition, the ***t*** and ***u*** vectors are derived such that they contain information about each other. This is explained in more detail below.

The ***t*** and ***u*** vectors are derived by applying the non-linear iterative partial least squares (NIPALS) algorithm [3]. The algorithm sequentially extracts a pair of ***t*** and ***u*** vectors, and then regresses them against each other. Then, it subtracts the regressed information from both the **X** and **Y** matrices, and proceeds to the extraction of the next pair of ***t*** and ***u*** vectors.

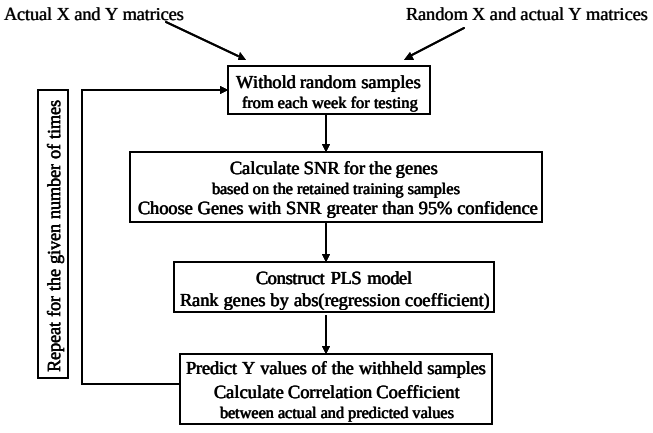

Fig. 1. Bootstrap PLS modeling methodology for obtaining consensus ranking of relevant genes. This procedure was repeated for 5000 trials.

The first $\boldsymbol{t}$ vector is the one that maximizes the information between it and the $\mathbf{Y}$ space. This is done by obtaining the first eigenvector $\boldsymbol{w}_1$ of the sample covariance matrix $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$, where the superscript T implies the transpose of the matrix. Then, the first input scores vector, $\boldsymbol{t}_1 = \mathbf{X}\boldsymbol{w}_1$, and the loading vector is obtained as $\boldsymbol{p}_1 = \mathbf{X}\,\boldsymbol{t}_1/\,\boldsymbol{t}_1{}^T\boldsymbol{t}_1$. The $\mathbf{Y}$ loading vector $\boldsymbol{q}_1$ is then determined by $\boldsymbol{q}_1 = \mathbf{Y}\,\boldsymbol{t}_1/\,\boldsymbol{t}_1{}^T\boldsymbol{t}_1$, and the output score vector $\boldsymbol{u}_1$ is obtained as $\boldsymbol{u}_1=\mathbf{Y}\boldsymbol{q}_1/\boldsymbol{q}_1{}^T\boldsymbol{q}_1$. The $\mathbf{X}$ and $\mathbf{Y}$ matrices are then deflated, and the process is repeated with the residual matrices. More details may be obtained in Geladi *et al*. (1986).

As the number of latent variables increases, the quality of regression improves, but so does the risk of over-fitting. A cross-validation procedure is implemented to circumvent this problem. One sample is withheld from the given data set, and the procedure described above is conducted to determine the latent variable at this stage. Using this latent variable (and any preceding ones), the values of the withheld samples are predicted, and the squared error between the actual and the predicted value recorded. This procedure is then repeated until all the samples have been withheld once, and the cumulative predictive error sum of squares (PRESS) is obtained. This is repeated at each stage, and finally the PRESS is plotted as a function of the latent variables. Dramatic increases in PRESS imply that over-fitting has occurred, and therefore, too many latent variables have been included. Typically, it is observed that PRESS initially declines as the number of LVs increase, and then rises. The number of latent variables chosen is the one that minimize the PRESS.

For our purposes, the $\mathbf{X}$ block data is the gene expression data, and the $\mathbf{Y}$ block data is the physiological variables measured, which in this case were insulin levels in the mice. Since microarray data displays a large amount of variability, a bootstrap methodology was implemented in order to develop robust classifiers and aid in the identification of a relevant set of variables or genes. The procedure consisted of withholding a certain number of samples for testing, and constructing the PLS model on the remaining samples. This involved pre-selecting genes based on their signal to noise ratio (SNR), and then choosing a PLS model based on cross validation. The model was then used to predict the value of the withheld samples, and the correlation coefficient between the predicted and actual values was calculated. Further, the genes were ranked by the absolute value of the regression coefficient. Then, a different set of samples was withheld, and the entire procedure repeated. The reason this repetitive procedure was implemented was due to the fact that a single PLS model is very dependent on the nature of the samples in the training data set. By repeating this entire procedure several thousand times, a more representative and valuable ranking of genes may be obtained. The consensus ranking of genes across these several thousand trials will be less dependent on any particular set of samples, and closer to the true biological significance. This process is illustrated in Figure 1.

III. RESULTS

A. *Experimental design and development of diabetes*

C57bl6 mice on a high fat diet develop diabetes, and the experiment was designed to track this development over various stages. The experiment design consisted of putting the C57 strain of mice on a high-fat and low-fat diet, and maintaining this diet for a period of 12 weeks. Power analysis was performed in order to determine the number of samples that needed to be used to observe reliable differential gene expression measurements. Based upon our analysis, 5 mice were sacrificed at the end of each week from each group, and the liver was extracted and stored. In addition, serum from each animal was extracted in order to make measurements on circulating insulin, and glucose levels. Before sacrificing, DEXA measurements were performed on the mice to determine the abdominal fat %. A total of 120 mice were used in this experiment, and for each mouse, the microarrays were conducted in duplicate, thus a total of 240 arrays were conducted. Physiological measurements revealed dramatic increases in insulin and abdominal fat % over the weeks for the C57 mice on a high fat diet, and are shown in Figure 2. The

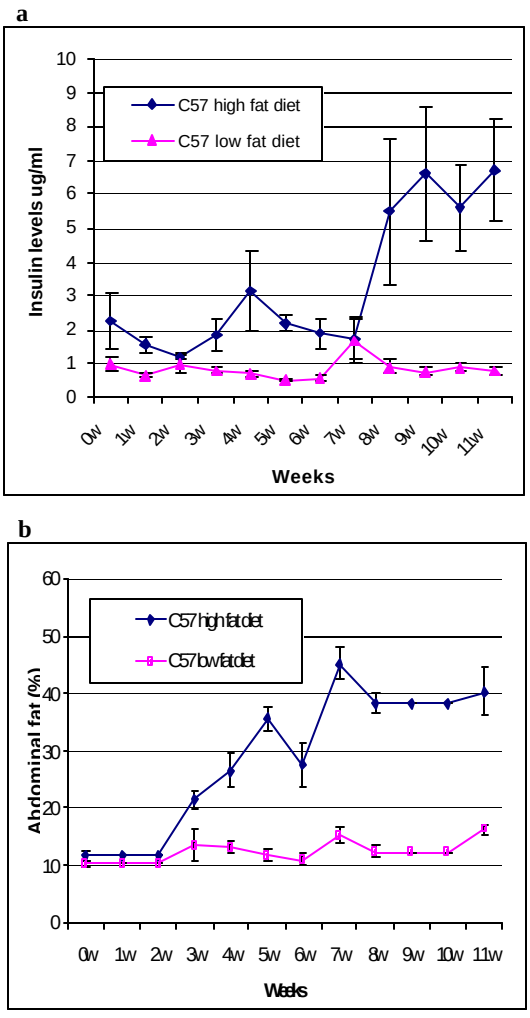

Fig. 2. Variation of (a) serum insulin levels and (b) the abdominal fat % for the C57 mice on a high fat and low fat diet.

glucose measurements were similar across the weeks, implying that the mice are controlling the level of glucose at the expense of insulin.

B. Model construction: Comparison of actual data and random data

In addition to the actual data, the bootstrap PLS method was also performed on a set of random data to act as a control on the model construction process. The correlation coefficient between the experimental Y (insulin) values and the predicted Y (insulin) values for both the transcriptional data and the random data were compared. It was found that the mean for the correlation coefficient for the random data was -0.002, while that for the transcriptional data was 0.67. The histogram of correlation coefficients is presented in Figure 3. The histograms illustrate that the expression measurements have a certain degree of predictive power that the random data set lacks. This is reassuring, since it increases our confidence in the data obtained through microarrays, and also the model construction procedure.

C. Good, average and bad PLS models

A comparison between a bad, a good and an average model is illustrative. Here, good, average, and bad is determined on the basis of the correlation coefficient between the predicted and the actual insulin levels. On comparing the lists of genes that were obtained for the three kinds of models with the genes that were obtained from the consensus rankings, it was found that the bad models had an extremely different set and ranking of genes, while both the good and the average models had similar rankings and content. The difference between the good and the average models were the samples that had been chosen in the training data set (Figure 4). Further, it was observed that the predicted values for the test set, in both the average and the good models had an upper limit, or a "glass ceiling". This is to say that the predicted insulin values based on just the gene expression levels had a maximum upper limit. This may be due to a variety of reasons. One of them may be that the

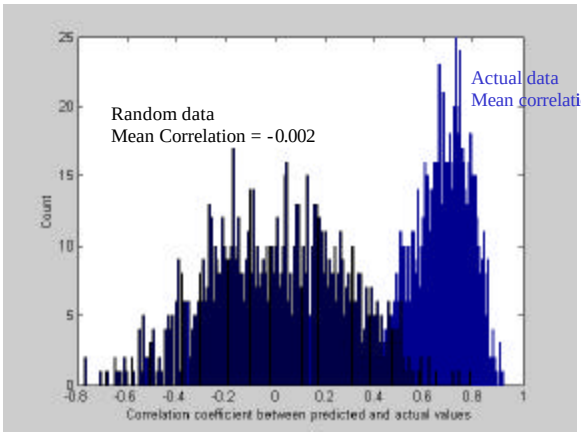

Fig. 3. Histogram of the correlation coefficient between the actual insulin levels and predicted insulin levels for the **test** data set in 5000 trials of the PLS model construction. Both random data and the actual gene expression data was used. The mean correlation coefficient for the random data was −0.002, and that for the actual data was much higher, about 0.67.

TABLE I
RANKED LIST OF GENES FROM PLS ANALYSIS

Gene ID	Description
BC002198	RIKEN cDNA 4833425P12 gene
BC003249	Expressed sequence AI646975
NM_013697	Transthyretin
NM_010560	Interleukin 6 signal transducer
NM_010358	Glutathione S-transferase, mu 1
NM_009481	Ubiquitin specific protease 9, X chromosome
AF349718	Procollagen, type IX, alpha 3
AK014145	RIKEN cDNA 3110038O15 gene
AK017001	RIKEN cDNA 4933431C10 gene
BC011111	RIKEN cDNA 0610038P07 gene
AK010783	RIKEN cDNA 2410127E18 gene
NM_013506	Eukaryotic translation initiation factor 4A2
BC006626	Expressed sequence C77440
AK020023	RIKEN cDNA 5830471E12 gene
AK003405	RIKEN cDNA 1110004B19 gene
NM_011664	Ubiquitin B
AK002477	RIKEN cDNA 0610010I06 gene
NM_009284	Signal transducer and activator of transcription 6
NM_007412	Adrenomedullin receptor
AK004835	RIKEN cDNA 1210002E11 gene
AK004636	SH3-domain kinase binding protein 1

relationship between gene expression measurements and the physiological data is non-linear, and therefore the linear modeling method is unable to provide a full description of the procedure.

D. Non-linear PLS

A non-linear PLS methodology based on the Implicit Non-Linear Regression (INLR) technique was implemented [4]. Non-linear PLS techniques are very flexible, and can tend to model noise. The INLR technique allows for the modeling of slight non-linearities, and is simple to implement. Instead of employing a non-linear inner relation, INLR squares the initial **X** variables to try and model the non-linearity. The rest of the PLS construction procedure remains identically similar.

It was found that after implementing a bootstrap non-linear PLS procedure, the mean correlation coefficient between the predicted and actual insulin values was 0.63 for actual data, while it was -0.075 for the random data matrix. The list of genes obtained was very similar to the consensus rankings obtained from the linear PLS. The decline in correlation coefficient to 0.63 from 0.67 may be an indication that the INLR technique is beginning to model noise as compared to linear PLS. Therefore, the non-linear approach was discontinued.

E. Discussion of the genes

The regression analysis provides a list of genes ranked by their regression coefficient in the bootstrap PLS. A partial list of the identified genes is provided in Table I Ranking gives an indication of the importance of the gene. The first gene in the PLS model (BC002198) is an unknown gene that has a 31%

amino acid homology with ankyrin 3. Ankyrins represent a protein family whose members are associated with membrane proteins and the actin cytoskeleton [5]. Liver Ankyrin 3 has not been associated with diabetes. The second gene in the PLS ranking (BC003249) is the Thyroid Hormone Receptor Interactor 10 a.k.a Trip10 or CDC42 interacting protein or CIP4/2. This gene was recently identified as a requirement for insulin-stimulated Glut-4 translocation in 3T3L1 mouse adipocyte cell lines [6]. Along the same pathway of function is another gene identified by PLS; SH3-domain kinase binding protein 1 (AK004636) a.k.a. CAP. Figure 5 shows the pathway and denotes the importance of those genes in the insulin mediated glucose reuptake.

Although only 3 genes are reported in this gene discussion, the full list of genes is currently being explored and many more pathways are being examined as relevant to the phenomenon of insulin resistance.

IV. DISCUSSION

This study attempted to develop quantitative, predictive models of physiology based on gene expression data. The results were encouraging, with a mean correlation coefficient of 0.67 between the actual and the predicted insulin measurements. The fact that only an average correlation coefficient of 0.67 between the predicted and actual insulin measurements for the test data set could be obtained, coupled with the fact of the observed "glass ceiling" effect would seem to imply that gene expression measurements, at least as regards predictive model construction, are semi-quantitative. This may be due to a variety of reasons:

- Quality of DNA microarray data. Given the large amount of variation observed in microarray data, it may not be possible to construct more predictive models. Improvements in DNA microarray technology can address this question.
- Observation of a limited set of data. Not all interactions occur at the transcriptional level, hence the data may be incomplete. Further, to completely characterize the circulating insulin levels, a systemic model of all the affected tissues may be needed.
- Biology may not function as a purely quantitative system, and a certain increase in the transcription of a gene may

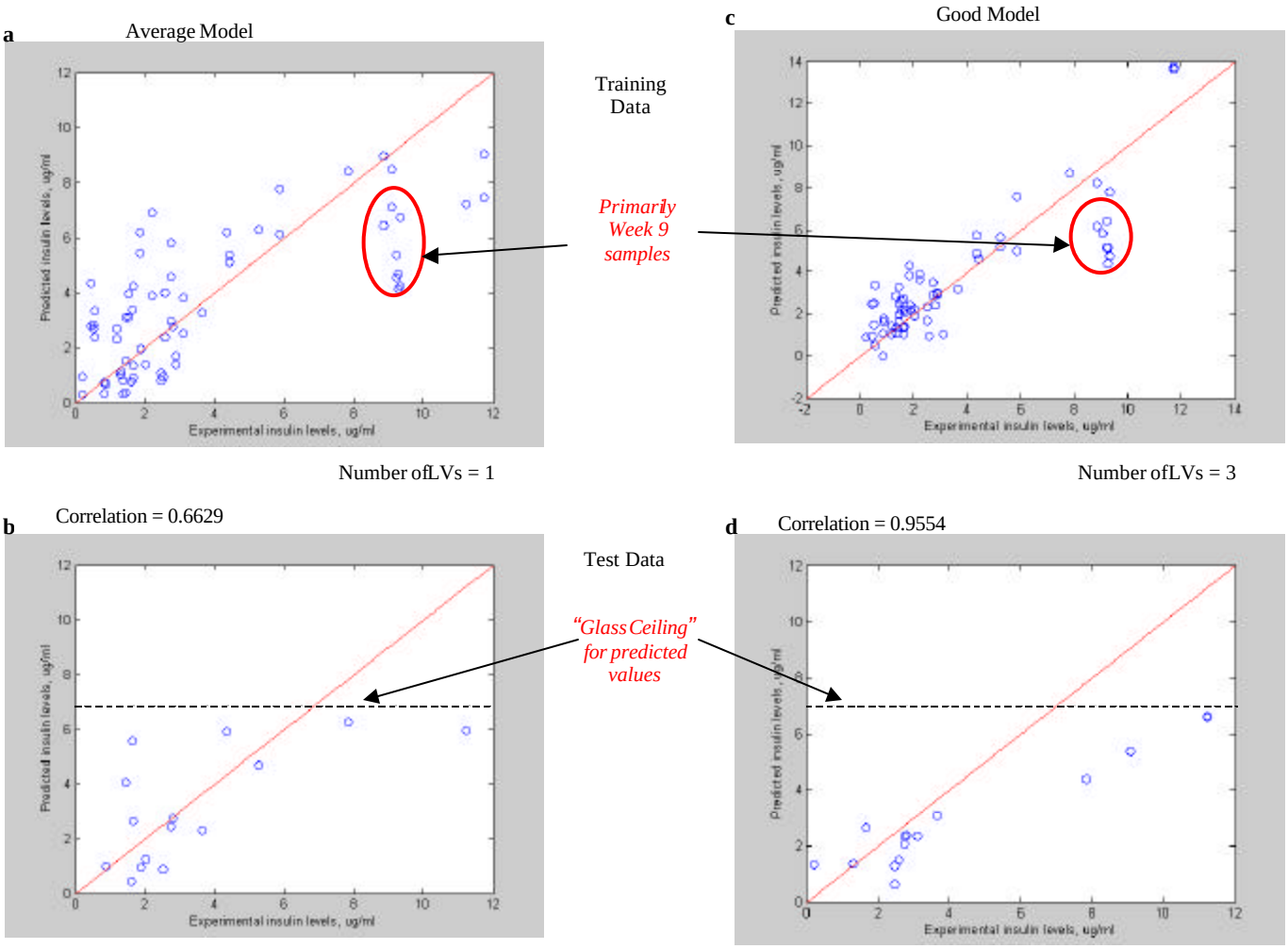

Fig. 4. Average and good PLS models. (a) and (b) are the training and test data performance for the average PLS model, while (c) and (d) are the training and test data performance for the good PLS model, respectively. All the graphs contain the predicted insulin levels on the y-axis in ug/ml, and the experimental insulin levels on the x-axis in ug/ml. In both cases, the samples measured in week 9 present difficulties in modeling. This is because the gene expression doesn't justify the larger increase in insulin for these samples. Also, in the case of predicting the test data, in parts (b) and (d), there is a upper limit in the predictions. Note that correlation values reported are for the test data set, not the training data set

not always yield the same response, due to a plethora of modulating effects.

Based on the histograms for the correlation coefficient, it may be seen that there is a fair degree of variation in the correlation coefficient for the test data. Therefore, construction of just a single PLS model is unjustified, and always a consensus building procedure such as the one implemented here must be employed.

An important aspect of the PLS based modeling methodology is that it can be expanded to include multiple sources and a variety of data, such as metabolic flux data, and protein expression data. In this way, it can allow for the construction of a more comprehensive predictive models for physiology.

Based on prior literature, several of the genes found to be important as a result of this analysis had already been implicated in the development of diabetes. This builds credence around the identified genes, and the model construction methodology.

The utility of viewing physiology as quantifiable and constructing predictive models based on just gene expression data is enormous. Potentially, the model allows for the construction of a space where different treatments may be evaluated on a common platform that does not rely on just a few variables. For example, the model of insulin resistance as developed here may be used to evaluate the efficacy of various treatments that retard this process in a quantifiable fashion by querying the model with the transcriptional profile of the treated samples. This approach, as opposed to just a simple measurement of insulin levels may provide more insight into the treatment chemical, since the evaluation is based on several gene expression measurements. Analogously, this may also lead to the development of diagnostic techniques that not only predict the occurrence/absence of a disease, but also if the disease is diagnosed, then to what extent has the disease progressed. The complete potential of these high throughput genomic methods in combination with sophisticated analytic tools can only be realized by conducting samples in a high-throughput fashion as well, specially given the large variance inherent in microarray data.

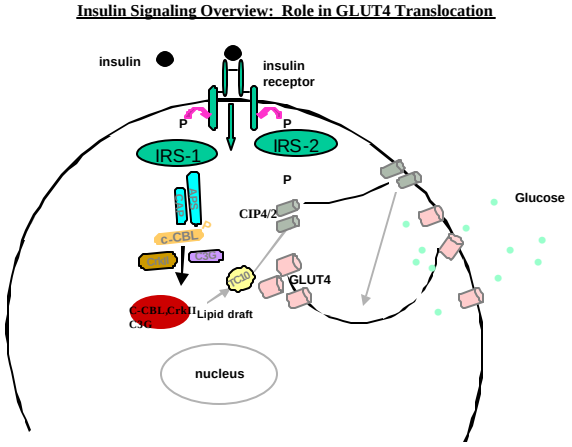

Fig. 5. CIP4/2 translocation to the plasma membrane is necessary for the GLUT4 membrane translocation. Both CIP4/2 and CAP were identified by our PLS model to be change their gene expression according to the concentration of circulating insulin.

ACKNOWLEDGMENT

The authors would like to thank Daehee Hwang, Bill Schmitt, and Michael Raab for helpful discussions regarding the modeling approach.

REFERENCES

[1] J.H. Kim, I.S. Kohane, L. Ohno-Machado, "Visualization and evaluation of clusters for exploratory analysis of gene expression data", *Journal of Biomedical Informatics*, vol. 35, pp. 25-36, February 2002.

[2] G. Stephanopoulos, D.H. Hwang, W.A. Schmitt, J. Misra, G. Stephanopoulos, "Mapping physiological states from microarray expression measurements", *Bioinformatics*, vol. 18, pp. 1054-1063, August 2002.

[3] P. Geladi and B.R. Kowalski, "Partial least squares regression: a tutorial", *Analytica Chimica Acta*, vol. 185, pp. 1-17, 1986.

[4] A. Berglund and S. Wold, "INLR, implicit non-linear latent variable regression", *Journal of Chemometrics*, vol.11, pp. 141-156, 1997.

[5] B. Peters, H.W. Kaiser, and T.M. Magin, "Skin-specific expression of ank-3(93), a novel ankyrin-3 splice variant", *Journal of Investigative Dermatology*, vol. 116, pp. 216-223, February 2001.

[6] L. Chang, R.D. Adams, A.R. Saltiel, "The TC10-interacting protein CIP4/2 is required for insulin-stimulated Glut4 translocation in 3T3L1 adipocytes", *Proceedings of the National Academy of Sciences*, vol. 99, pp. 12835-12840, October 2002.