Improved prediction of protein interaction from microarray data using asymmetric correlation

Kojiro Yano*[1]

[1]Department of Physiology, Development and Neuroscience,University of Cambridge, Downing Street, Cambridge, UK

Email: Kojiro Yano*- ky231@cam.ac.uk;

*Corresponding author

Nature Precedings : hdl:10101/npre.2010.4430.1 : Posted 9 May 2010

Abstract

Background: Detection of correlated gene expression is a fundamental process in the characterization of gene functions using microarray data. Commonly used methods such as the Pearson correlation can detect only a fraction of interactions between genes or their products. However, the performance of correlation analysis can be significantly improved either by providing additional biological information or by combining correlation with other techniques that can extract various mathematical or statistical properties of gene expression from microarray data. In this article, I will test the performance of three correlation methods-the Pearson correlation, the rank (Spearman) correlation, and the Mutual Information approach-in detection of protein-protein interactions, and I will further examine the properties of these techniques when they are used together. I will also develop a new correlation measure which can be used with other measures to improve predictive power.

Results: Using data from 5,896 microarray hybridizations, the three measures were obtained for 30,499 known protein-interacting pairs in the Human Protein Reference Database (HPRD). Pearson correlation showed the best sensitivity (0.305) but the three measures showed similar specificity (0.240 - 0.257). When the three measures were compared, it was found that better specificity could be obtained at a high Pearson coefficient combined with a low Spearman coefficient or Mutual Information. Using a toy model of two gene interactions, I found that such measure combinations were most likely to exist at stronger curvature. I therefore introduced a new measure, termed asymmetric correlation (AC), which directly quantifies the degree of curvature in the expression levels of two genes as a degree of asymmetry. I found that AC performed better than the other measures, particularly when high specificity was required. Moreover, a combination of AC with other measures significantly improved specificity and sensitivity, by up to 50%.

Conclusions: A combination of correlation measures, particularly AC and Pearson correlation, can improve prediction of protein-protein interactions. Further studies are required to assess the biological significance of asymmetry in expression patterns of gene pairs.

Background

In microarray data analysis, it is common to examine correlations among gene expression levels [1]. Correlated expression of a group of genes implies that the genes are involved in the same biological process or form a protein complex, and correlation measures have been used to predict disease markers [2] and protein-interaction partners [3]. For linear correlations, measures such as Euclidean distance or Pearson correlation (PC) have been used, whereas more general correlations may be quantified by rank (Spearman) correlation (SC) [4], Mutual Information (MI) [5]. However, the reliability of these measures when used to infer gene interactions is not always satisfactory [6], and new methods have been proposed to improve the functional annotation of genes from microarray data. Some of these techniques seek to enhance the performance of measures by incorporating evolutionary information, such as orthologous co-expression [7, 8], or by considering conditional correlations mediated by a third gene [9–11]. In addition to correlation measures, statistically more sophisticated but computationally more intensive methods have also been developed; these include Bayesian networks [12–14] and support vector machines [15, 16].

Gene interactions cover a wide variety of mechanisms, and different methods of gene network inference are based on various models of gene interaction. Therefore, a combination of such approaches can improve the performance of network inference. For example, it has been shown that direct inference methods such as PC are suitable for detecting stable protein complexes whereas conditional methods, including partial PC or the Graphical Gaussian model, are better at defining causal interactions [17]. This means that different methods can be mutually complementary, when used to expand detection of protein interactions. In this article, I examine the performance of three commonly used measures - PC, SC, and MI - in predicting known human protein-protein interactions from microarray data, and assess the possibility of achieving improved performance through data combination. Based on this analysis, I introduce a new measure, termed asymmetric correlation (AC), and show that AC improves the performance of other measures.

Results and Discussion

Data source

The pre-processed meta-analysis data set E-TABM-185 from ArrayExpress [18] was used in all analyses. The dataset contained measurements from 5,896 arrays of human cell and tissue samples, all of which were hybridized with the Affymetrix GeneChip Human Genome HG-U133 (22215 probes). In this dataset, raw data from the microarray were re-normalized with *gcrma* from Bioconductor [19] and output data were all log2-transformed. Human protein interaction data were from the Human Protein Reference Database (HPRD) [20], which has information on 30,499 interactions (excluding self-interactions). I also randomly selected 30,499 pairs of genes from the microarray dataset, as a negative control.

Quantification of correlated gene expression

PC and SC were calculated using *corr* function from MATLAB (Mathworks, Inc. Natick, MA) and MI was estimated using *information* MATLAB function constructed by Moddemeijer [5]. More specifically, when I(X;Y) is MI of variables X and Y, $I(X;Y) = H(X) + H(Y) - H(X,Y)$, where H(X) are H(Y) are the entropies of variable X and Y, respectively and H(X,Y) is the joint entropy of X and Y. To obtain H(X), the maximum and minimum values of X, namely X_{max} and X_{min} were calculated and the range $X_{min} \leq X \leq X_{max}$ were divided into ten equal intervals (called bins) $B_i = B_1$, B_2..., B_{10}. Next the number of elements C_i in bin B_i was calculated for all bins. Finally H(X) was calculated as $H(X) = -\sum P_i(X)log(P_i(X))$, where $P_i(X)$=$C_i/\sum C_i$. H(Y) is obtained by the same manipulation. To obtain H(X,Y), the number of elements $C_{i,j}$ of a vector $Z_{i,j}$=(X_i, Y_i) was calculated for 10-by-10 bins $B_{i,j}$ with equal intervals in the ranges $X_{min} \leq X \leq X_{max}$ and $Y_{min} \leq Y \leq Y_{max}$. Next H(X,Y) was calculated as $H(X,Y) = -\sum P_{i,j}(X,Y)log(P_{i,j}(X,Y))$, where $P_{i,j}(X,Y)$=$C_{i,j}/\sum C_{i,j}$. Correlation measures for both protein-protein interaction pairs and control pairs are shown as histograms in Figure 1. With all three measures, mean values were higher using HPRD genes than control genes (*HPRD*; PC, 0.167; SC, 0.244; MI, 0.0981; *Control*: PC, 0.0139; SC, 0.1127; MI, 0.0558).

Sensitivity and specificity of protein interaction prediction by gene expression correlation

To compare the performance of the three measures, I calculated sensitivity and specificity in discovery of protein-protein interactions. At various threshold levels above which correlations were considered to be significant, specificity (proportion of below-threshold pairs to the total in the control group) and sensitivity (proportion of above-threshold pairs to the total in the protein-interaction group) were calculated, and are plotted in Figure 2. When the specificity was 0.9, the sensitivity was 0.305 with PC, 0.255 with MI, and 0.125 with SC. When the sensitivity was 0.9, the specificity was 0.257 with PC, 0.244 with MI, and 0.240 with SC. Therefore, PC performed best when high specificity was required, but there was no significant difference in measure performance at high sensitivity.

Combined measures yield better specificity

To improve prediction performance, I compared the combined distributions of pairs of measures using the HPRD pairs and controls. Figure 3 shows scatter plots for PC and SC, PC and MI, and SC and MI, for all pairs in the protein-interacting group (blue) and the control group (red). It was found that at high PC (>0.5), protein-interacting pairs became more predominant in the lower range of MI or SC. Such properties were not seen when the relationship between SC and MI was examined; in all ranges of SC the protein-interacting group predominated over the control group at higher levels of MI.

In what situation will SC or MI be reduced when PC is still high? I approach this question using a toy model. Let us consider genes X and Y and assume that the expression level of gene Y (termed Y_e) is a function of the expression level of gene X (termed X_e), and that they system follows Michaelis-Menten kinetics, as shown below:

$$dY_e/dt = V_1 X_e/(X_e + K) - k_e Y_e$$

where V_1, K and k_e are constants. At the steady state, the relationship between Y_e and X_e is:

$$Y_e = V_1 X_e/k_e/(X_e + K) = V_2 X_e/(X_e + K)$$

where $V_2 = V_1/k_e$. Now the measured levels of Y_e and X_e (termed Y_m and X_m) are expressed as

$$Y_m = d_1 Y_e + D_1, X_m = d_2 X_e + D_2$$

where d_1 and d_2 are biological noise modelled as normally distributed noise with a mean = 1 and a standard deviation = R_1. D_1 and D_2 are technical noise modelled as evenly distributed random variables ranging between 0 and R_2 [21]. Note that the parameters for noise are arbitrarily chosen and are used solely for demonstration of potential effects on correlation measures. Now, I vary K, R_1, and R_2 as follows: $K = 10,000, 1,000$ or 100, and $[R_1, R_2] = [0.01, 10], [0.03, 30]$ or $[0.09, 90]$. Figure 4 shows the log2 transformed plot of $X_m (10 < X_m < 1,000)$ against Y_m and Table 1 shows PC, SC and MI outputs with different combinations of K and $[R_1, R_2]$. At low and moderate noise levels ($[R_1, R_2] = [0.01, 10]$ and $[0.03, 30]$, respectively), SC was relatively unaffected by the decrease in K (i.e. the increase of non-linearity). At the high noise level ($[R_1, R_2] = [0.09, 90]$), however, SC was strongly affected by the decrease in K, but PC was less affected. At all noise levels, MI was highly sensitive to the value of K. These results show that the elevated PC, and the low SC and MI, on analysis of the protein-interacting group, are more likely to be evident at high levels of non-linearity and noise. As the level of noise will not be significantly different between HPRD and control groups, nonlinearity will probably be higher when the protein-interacting pairs are considered.

Defining asymmetric correlation

In the toy model above, the increase in non-linearity seen as K decreased caused the plot to diverge from the diagonal line and become more asymmetric. This has been noted not only when Michaelis-Menten kinetic is applied, but also in a system where reaction activation involves cooperativity between activators [22–24]. I now attempt to define and quantify the asymmetry of expression of a pair of genes. When the expression levels of genes X and Y are X_e and Y_e, and their means are X_{av} and Y_{av}, respectively, I define

$$Q = \sum |(X_e - X_{av})(Y_e - Y_{av})|$$

and

$$Q_1 = Q \text{ for all } (X_e,Y_e) \text{ with } X_e > X_{av} \ \& \ Y_e > Y_{av},$$

$$Q_2 = Q \text{ for all } (X_e,Y_e) \text{ with } X_e < X_{av} \ \& \ Y_e > Y_{av},$$

$$Q_3 = Q \text{ for all } (X_e,Y_e) \text{ with } X_e < X_{av} \ \& \ Y_e < Y_{av},$$

$$Q_4 = Q \text{ for all } (X_e,Y_e) \text{ with } X_e > X_{av} \ \& \ Y_e < Y_{av},$$

X_e and Y_e are symmetric when all of Q_1, Q_2, Q_3 and Q_4 are equal, and asymmetric if they are not. Please note that linear plots, such as Y=X are also considered to be asymmetric by this definition. Therefore nonlinear curves are asymmetric but asymmetry alone does not guarantee nonlinearity.

Now let me give a simple example to explain this concept more clearly. Here is a microarray dataset Z which has three samples Z_1, Z_2 and Z_3, with expression levels of genes X and Y being (X_1,Y_1), (X_2,Y_2) and (X_3,Y_3), respectively, and X_1=5,X_2=X_3=2, Y_1=Y_2=5, Y_3=2, Z_{av} = (X_{av}, Y_{av})=(3,4) (Figure 5). Therefore, according to the definition above, $Q_1 = |(X_1 - X_{av})(Y_1 - Y_{av})|$=2, which is the area of the rectangle defined by Z_1 and Z_{av}. Similarly, $Q_2 = |(X_2 - X_{av})(Y_2 - Y_{av})|$=1, $Q_3 =$ $|(X_3 - X_{av})(Y_3 - Y_{av})|$=2 and Q_4= 0. Therefore Z is asymmetric.

When Q_z is the smallest value of Q_1...Q_4, X_e and Y_e are termed asymmetric with respect to Q_z. Alternatively, it could be stated that X_e and Y_e are asymmetric *with respect to the upper right quadrant* if Q_1 is the smallest, *with respect to the upper left quadrant* if Q_2 is the smallest, *with respect to the lower left quadrant* if Q_3 is the smallest, or *with respect to the lower right quadrant* if Q_4 is the smallest. Finally *the asymmetric correlation coefficient* S_z with respect to Q_z is defined by $S_z = \sum Q_i/4 - Q_z$. In the example above, Z is asymmetric with respect to Q_4, or lower right quadrant, and the asymmetric coefficient S_4 = (Q_1+Q_2+Q_3 + Q_4)/4 - Q_4= ($|(X_1 - X_{av})(Y_1 - Y_{av})|$+$|(X_2 - X_{av})(Y_2 - Y_{av})|$+$|(X_3 - X_{av})(Y_3 - Y_{av})|$)/4=(2+1+2)=1.25. On the other hand, if X_2=(X_1+X_3)/2 = 3.5 and Y_2=(Y_1+Y_3)/2=3.5, Z_1, Z_2 and Z_3 will form a straight line and $S_4 =$ ($|(5 - 3.5)(5 - 3.5)|$+0+$|(2 - 3.5)(2 - 3.5)|$)/4=1.125, which is smaller than when Z was asymmetric.

Inplementation of the asymmetric correlation coefficient

A difficulty with asymmetric correlation is that it is highly sensitive to the distribution of X_e and Y_e. In other words, X_e and Y_e will be asymmetric if their distribution is skewed, even when they are entirely independent. Therefore it is important to eliminate the asymmetry which comes from the individual distribution of X_e and Y_e. I employed the following procedures to this end. First, an $n * n$ two-dimensional matrix $M_{s_{x,y}}$ is calculated from a two-dimensional histogram of a scatter plot for $V(X_e, Y_e)$, where V is an expression level vector for genes X and Y. Next two $1 * n$ one-dimensional vectors V_X and V_Y are calculated from one-dimensional histograms of X_e and Y_e, respectively. In the third step a matrix $M_{r_{x,y}}$ is derived by $M_{r_{x,y}} = V_X^{'} * V_Y/$(total number of samples). This matrix represents the expected distribution of combinations of X_e and Y_e when X_e and Y_e are independent. Finally, I normalize $M_{s_{x,y}}$ by $M_{r_{x,y}}$ to obtain the normalized expression matrix $M_{n_{x,y}} = M_{s_{x,y}}./(M_{s_{x,y}} + M_{r_{x,y}})$, where ./ indicates element-wise division. Now, the asymmetric coefficient with respect to the lower right quadrant S_4 is obtained by $\sum_{i=1}^{4} S_i/4 - \sum S_4 =$ $\sum_{x=1,y=1}^{n,n}(M_{n_{x,y}}|(x-(n+1)/2)(y-(n+1)/2)|)/4 - \sum_{x=n/2,y=1}^{n,n/2}(M_{n_{x,y}}(x-(n+1)/2)((n+1)/2-y))$.

Performance of asymmetric correlation

Figure 6 shows histograms of the asymmetric correlation coefficient with respect to the lower right quadrant for HPRD data and the control data used above. The mean values of the coefficients are 1,599 for

HPRD data and 1,002 for control data. When specificity and sensitivity were examined as for other correlation measures, the asymmetric coefficient performed better than did PC. When the specificity was 0.9, sensitivity was 0.347 (0.305 with PC), and when the sensitivity was 0.9, specificity was 0.290 (0.257 with PC).

However, better performance could be obtained by combining the asymmetric coefficient with other measures. Figure 7 shows scatter plots for asymmetric coefficients and PC, SC, and MI of all HPRD pairs (blue) and control pairs (red). As can be seen in the top-right portion of all three scatter plots, the HPRD pairs were very specifically detected when AC was high and the other measure (i.e. PC, SC or MI) was positive. As a result of this observation, I introduced a measure which combines PC, SC or MI with AC:

$$R_{PC,AC} = rP + A$$

$$R_{SC,AC} = rS + A$$

$$R_{MI,AC} = rM + A$$

where $R_{PC,AC}$, $R_{SC,AC}$ and $R_{MI,AC}$ are combined measures for PC and AC, SC and AC, and MI and AC, respectively. P, S, M and A in the equations are values of PC, SC, MI, and AC respectively and r is a constant. Figure 8 shows specificity at 90% sensitivity and sensitivity at 90% specificity when r was varied between 0 and 5,000. With PC and MI, specificity was relatively unaffected by r (maximum values = 0.293 and 0.290 for PC and MI, respectively). However, the sensitivity increased significantly as r increased and attained peaks (0.448 for PC and 0.389 for MI) when r = 2,400 and 5,000, respectively. On the other hand, both specificity and sensitivity improved with SC when r was increased to some extent, but all improvements were modest (peak specificity and sensitivity = 0.378 and 0.313, respectively, at r = 1,000) and a further increase in r caused significant deterioration in specificity. These results demonstrate that a combination of asymmetric correlation with other measures could improve performance by as much as 50%, depending on the measures chosen for combination.

Why did AC perform better when it was combined with PC? Both PC and AC can show large magnitudes for both linear and non-linear curves, but PC favors linear curves and AC favors non-linear curves. Therefore non-linear curves can be captured better by selecting gene pairs with relatively high AC and low PC. In Figure 7, some protein-interacting pairs which did not overlap with the negative controls had high PC (¿0.8) and modest AC (2000 2500), meaning they had linear relationships. However, there are also protein-interacting pairs which had high AC (¿3000) with relatively low PC (0.4), corresponding to nonlinear asymmetric relationships, and they were found more frequently from protein-interacting pairs than the negative controls. Therefore the combination of PC and AC is more robust than using PC or AC alone in detecting both linear and nonlinear relationships between protein-interacting pairs.

Conclusions

A number of sophisticated mathematical and computational microarray analysis techniques have been developed in recent years, but more work is needed to exploit the wealth of gene expression data in microarray and other databases. I found that, by using a new asymmetric correlation measure, it was possible to extract new information from passive microarray data (i.e. with no external perturbations) at a relatively low computational cost. However, the proposed method is not intended to replace existing methods for gene network inference, because gene regulatory mechanisms are so complex that no single analysis can ever extract all information from microarray data. Indeed, I found that asymmetric correlation can enhance the performance of existing inference methods when the techniques are used together. However, it should be noted that the method will not work when microarray data show little nonlinearity, in which case the PC and MI measures will demonstrate linear relationships. Moreover, asymmetry is just one of the many properties of nonlinearity and even the combination of AC and PC will not quantify the all properties of nonlinearity. Although I do not fully explore the topic in this article, another important property of asymmetric correlation is that it can produce directed graphs based on the direction of asymmetry. As asymmetry is one of the fundamental properties of causal relationships [25, 26], it may be possible to extract information on causality from asymmetric correlations. This should be examined systematically in future work, using both simulated datasets and a large collection of experimental data on causal relationships between proteins.

Author contributions

All work was carried out by K.Y.

Acknowledgements

K.Y. is the AstraZeneca Senior Research Fellow in Systems Biology, at Pembroke College, University of Cambridge, UK.

References

1. Baxevanis AD, Ouellette BFF (Eds): *Bioinformatics : a practical guide to the analysis of genes and proteins.* John Wiley & Sons, Inc., 3rd edition 2005.
2. Dupuy A, Simon RM: **Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting.** *J Natl Cancer Inst* 2007, **99**(2):147–157.
3. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions.** *Genome Res* 2002, **12**:37–46.
4. Zou KH, Tuncali K, Silverman SG: **Correlation and simple linear regression.** *Radiology* 2003, **227**(3):617–622.
5. Moddemeijer R: **On Estimation of Entropy and Mutual Information of Continuous Distributions**. *Signal Processing* 1989, **16**(3):233–246.
6. Almudevar A, Klebanov LB, Qiu X, Salzman P, Yakovlev AY: **Utility of correlation measures in analysis of gene expression.** *NeuroRx* 2006, **3**(3):384–395.
7. Tirosh I, Barkai N: **Computational verification of protein-protein interactions by orthologous co-expression.** *BMC Bioinformatics* 2005, **6**:40.
8. Fraser HB, Hirsh AE, Wall DP, Eisen MB: **Coevolution of gene expression among interacting proteins.** *Proc Natl Acad Sci U S A* 2004, **101**(24):9033–9038.
9. Soong TT, Wrzeszczynski K, Rost B: **Physical protein-protein interactions predicted from microarrays**. *Bioinformatics* 2008, **24**:2608.
10. Zhang J, Ji Y, Zhang L: **Extracting three-way gene interactions from microarray data.** *Bioinformatics* 2007, **23**(21):2903–2909.
11. Luo W, Hankenson KD, Woolf PJ: **Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information.** *BMC Bioinformatics* 2008, **9**:467.
12. Pe'er D, Regev A, Elidan G, Friedman N: **Inferring subnetworks from perturbed expression profiles.** *Bioinformatics* 2001, **17 Suppl 1**:S215–S224.
13. Djebbari A, Quackenbush J: **Seeded Bayesian Networks: Constructing genetic networks from microarray data**. *BMC Systems Biology* 2008, **2**:57.
14. Werhli AV, Grzegorczyk M, Husmeier D: **Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks**. *Bioinformatics* 2006, **22**:2523–2531.
15. Zhang C, Li P, Rajendran A, Deng Y, Chen D: **Parallelization of multicategory support vector machines (PMC-SVM) for classifying microarray data.** *BMC Bioinformatics* 2006, **7 Suppl 4**:S15.
16. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: **Multiclass cancer diagnosis using tumor gene expression signatures.** *Proc Natl Acad Sci U S A* 2001, **98**(26):15149–15154.
17. Zampieri M, Soranzo N, Altafini C: **Discerning static and causal interactions in genome-wide reverse engineering problems.** *Bioinformatics* 2008, **24**(13):1510–1515.
18. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A: **ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression.** *Nucleic Acids Res* 2009, **37**(Database issue):D868–D872.
19. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**(10):R80.

20. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Kishore CJH, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A: **Human Protein Reference Database–2009 update.** *Nucleic Acids Res* 2009, **37**(Database issue):D767–D772.

21. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data.** *Bioinformatics* 2002, **18**(11):1454–1461.

22. Benecke A: **Genomic Plasticity and Information Processing by Transcription Coregulators**. *Complexus* 2003, **1**:65.

23. Bolouri H, Davidson EH: **Modeling transcriptional regulatory networks**. *Bioessays* 2002, **24**:1118–1129.

24. Veitia RA: **A sigmoidal transcriptional response: cooperativity, synergy and dosage effects.** *Biol Rev Camb Philos Soc* 2003, **78**:149–170.

25. Hausman D: *Causal Asymmetries.* Cambridge: Cambridge University Press 1998.

26. Shipley B: *Cause and Correlation in Biology.* Cambridge: Cambridge University Press 2002.

Figures

Figure 1. Distribution of correlation measures for the Human Protein Reference Database (HPRD) and control data.

Pearson coefficient, Spearman coefficient, and Mutual Information were calculated for 30,499 known interacting pairs from HPRD (top) and the same number of randomly selected negative control pairs (bottom).

Figure 2. Sensitivity and specificity for detection of protein-protein interactions, and in distinguishing such interactions from random pairs.

Pearson coefficient (black), Spearman coefficient (blue), and Mutual Information (red) were calculated for gene pairs from the protein-protein interaction data in the Human Protein Reference Database (HPRD) and randomly selected negative control pairs. At different threshold levels, the proportion of HPRD pairs with above-threshold values, with respect to the total (sensitivity), and the proportion of control pairs with below-threshold values, with respect to the total (specificity), were calculated.

Figure 3. Scatter plots of correlation measures for the Human Protein Reference Database (HPRD) and control data.

Pearson coefficient, Spearman coefficient, and Mutual Information for known interacting pairs from HPRD (blue) and randomly selected control pairs (red) were plotted against each other. *left*, Pearson coefficient against Spearman coefficient; *Middle*, Pearson coefficient against Mutual Information; *Right*, Spearman coefficient against Mutual Information.

Figure 4. Measured expression levels of genes X and Y in the toy model at different levels of nonlinearity and noise

The scatter plots show log2-transfored measured expression levels of genes X and Y in the toy model. Nonlinearity (K) and noise levels (R_1 and R_2) were varied between 100 and 1000, 0.01 and 0.09, and 10 and 90, respectively, and their effects on the distribution of the measured levels of X and Y are shown.

Figure 5. An example for quantification of asymmetric correlation

This is a scatter plot of the expression levels of genes X and Y (X_e and Y_e, respectively) for microarray data Z with three samples, Z_1, Z_2 and Z_3 (open circles). Z_{av} (filled circle) is the average of Z_1, Z_2 and Z_3. The asymmetric coefficient Q is calculated from Q_1, Q_2, Q_3 and Q_4 and the area of the hatched rectangle in the figure represent Q_1, Q_2, and Q_3 as indicated. Q_4 is zero in this example.

Figure 6 - Distribution of asymmetric correlation for the Human Protein Reference Database (HPRD) and the control datasets.

Asymmetric correlation with respect to the lower-right quadrant (please see the main text for definition) was calculated for protein-interacting pairs from the HPRD (top) and negative control pairs from the control (bottom) datasets.

Figure 7 - Scatter plots of asymmetric and Pearsson coefficients for the Human Protein Reference Database (HPRD) and control datasets.

Asymmetric coefficient and Pearson coefficient (left), Spearman coefficient (center) and Mutual Information (right) for known interacting pairs from HPRD (blue) and randomely selected control pairs (red) were plotted against each other. Note the top right-hand corner of the scatter plots where blue spots are predominant.

Figure 8 - Sensitivity and specificity of a combined measure for detecting protein-protein interactions

The performances of a combined measure R=r(Pearson coefficient, Spearman coefficient or Mutual Information)+(asymmetric coefficient) are shown. The plot shows the proportion of HPRD pairs with above-threshold values with respect to total pairs (sensitivity, shown in blue) at 90% specificity, and the proportion of control pairs with below-threshold values with respect to total pairs (specificity, shown in red) at 90% sensitivity, respectively. Solid lines with no markers indicate values obtained with a combination of the asymmetric coefficient and the Pearson coefficient, those with square markers indicate values obtained with a combination of the asymmetric coefficient and the Spearman coefficient, and those with circles indicate values obtained with a combination of the asymmetric coefficient and Mutual Information.

Tables

Table 1. Correlation measures for measured expression levels of genes X and Y in the toy model using different combinations of parameters.

Parameters	Pearson coefficient	Spearman coefficient	Mutual Information
$K = 10000$, $[R_1, R_2] = [0.01, 10]$	0.999	0.999	1.33
$K = 1000$, $[R_1, R_1] = [0.01, 10]$	0.993	0.999	1.17
$K = 100$, $[R_1, R_2] = [0.01, 10]$	0.932	0.989	0.768
$K = 10000$, $[R_1, R_2] = [0.03, 30]$	0.994	0.995	1.31
$K = 1000$, $[R_1, R_2] = [0.03, 30]$	0.985	0.993	1.19
$K = 100$, $[R_1, R_2] = [0.03, 30]$	0.918	0.942	0.793
$K = 10000$, $[R_1, R_2] = [0.09, 90]$	0.957	0.962	0.960
$K = 1000$, $[R_1, R_2] = [0.09, 90]$	0.955	0.947	0.907
$K = 100$, $[R_1, R_2] = [0.09, 90]$	0.845	0.776	0.564

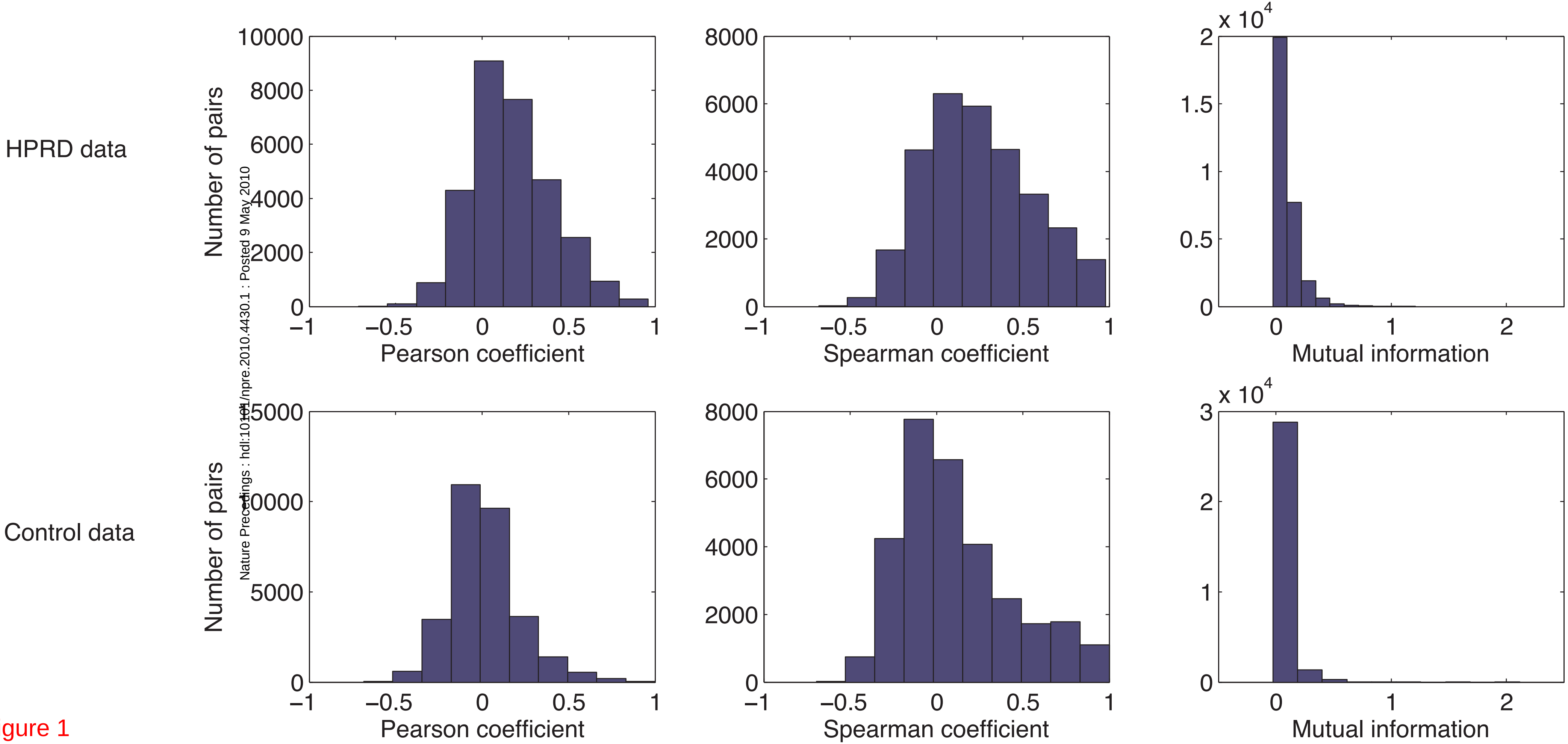

HPRD data
Number of pairs
10000
8000
6000
4000
2000
0
−1
−0.5
0
0.5
1
Pearson coefficient
8000
6000
4000
2000
0
−1
−0.5
0
0.5
1
Spearman coefficient
x 10^4
2
1.5
1
0.5
0
0
1
2
Mutual information
Control data
Number of pairs
15000
10000
5000
0
−1
−0.5
0
0.5
1
Pearson coefficient
8000
6000
4000
2000
0
−1
−0.5
0
0.5
1
Spearman coefficient
x 10^4
3
2
1
0
0
1
2
Mutual information

Figure 1

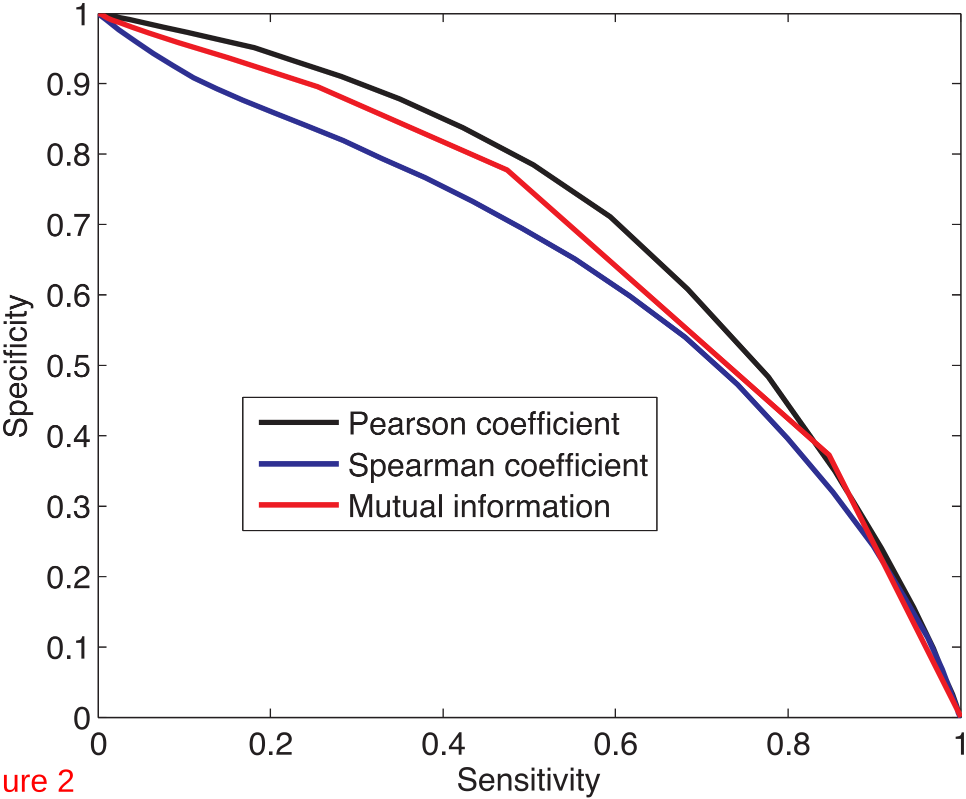

Specificity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
Sensitivity
Pearson coefficient
Spearman coefficient
Mutual information

Figure 2

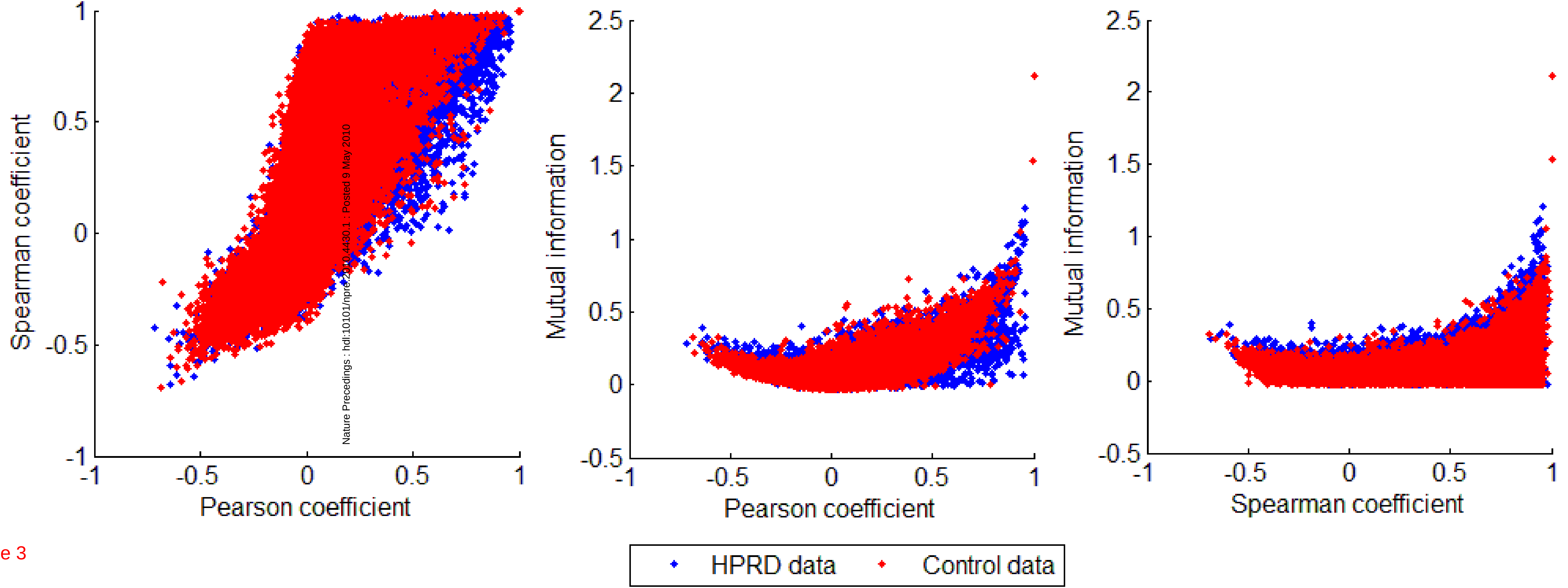

1
0.5
0
-0.5
-1
Spearman coefficient
-1
-0.5
0
0.5
1
Pearson coefficient
2.5
2
1.5
1
0.5
0
-0.5
Mutual information
-1
-0.5
0
0.5
1
Pearson coefficient
2.5
2
1.5
1
0.5
0
-0.5
Mutual information
-1
-0.5
0
0.5
1
Spearman coefficient
HPRD data
Control data

Figure 3

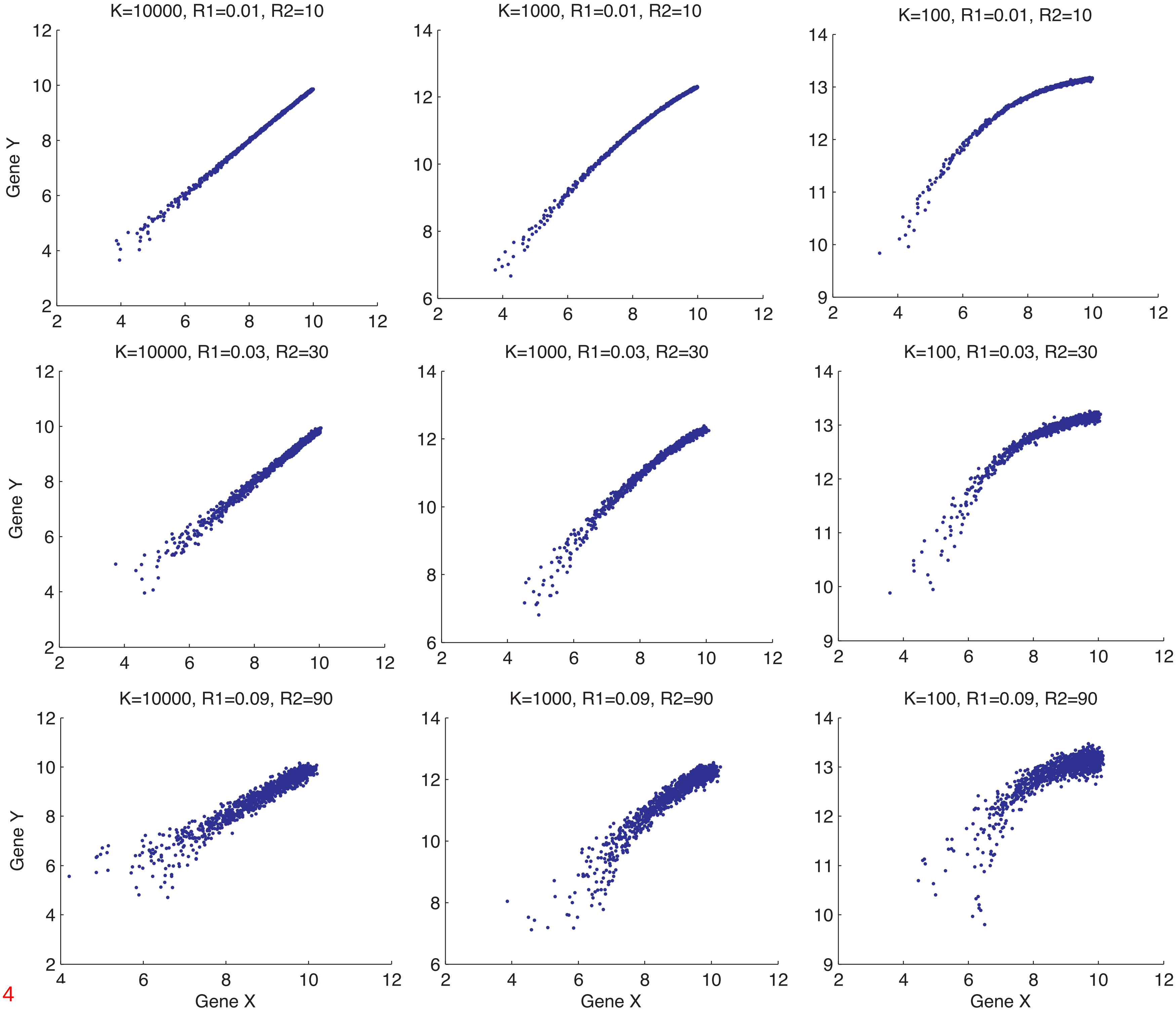

K=10000, R1=0.01, R2=10
K=1000, R1=0.01, R2=10
K=100, R1=0.01, R2=10
K=10000, R1=0.03, R2=30
K=1000, R1=0.03, R2=30
K=100, R1=0.03, R2=30
K=10000, R1=0.09, R2=90
K=1000, R1=0.09, R2=90
K=100, R1=0.09, R2=90
Gene Y
Gene Y
Gene Y
Gene X
Gene X
Gene X
2
4
6
8
10
12
14
9
11
13

Figure 4

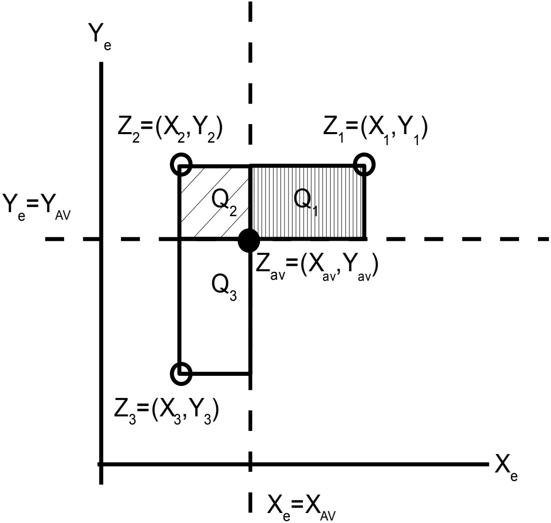


Figure 5

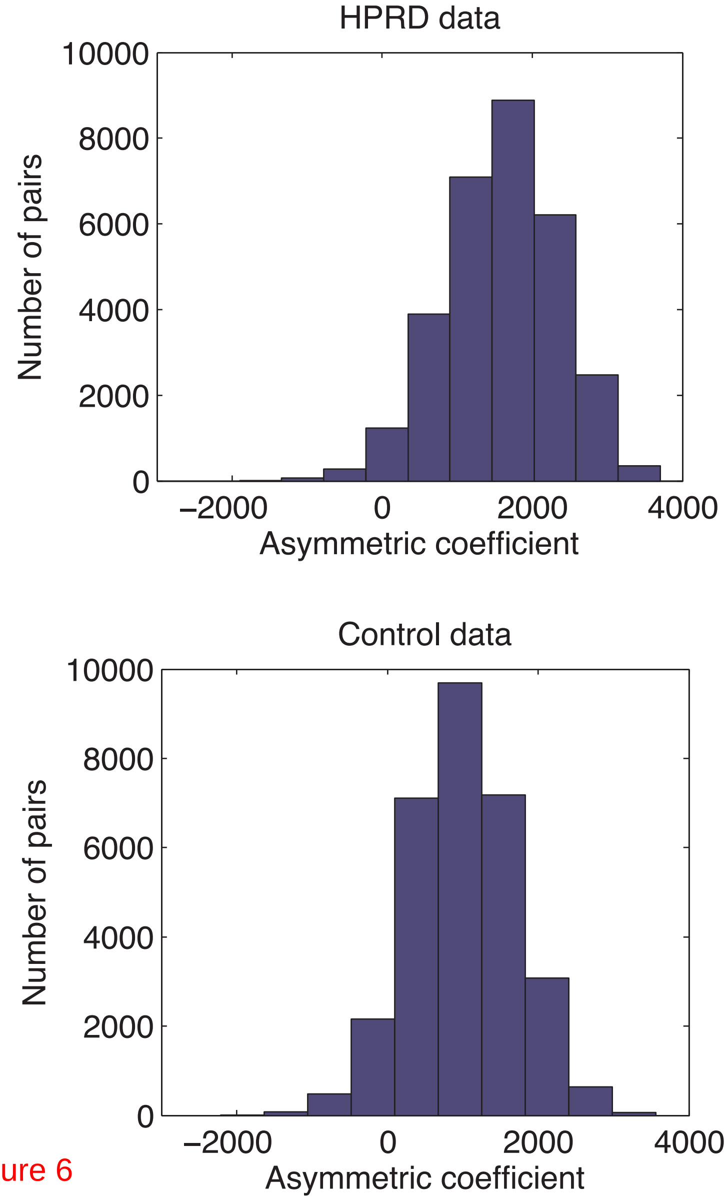
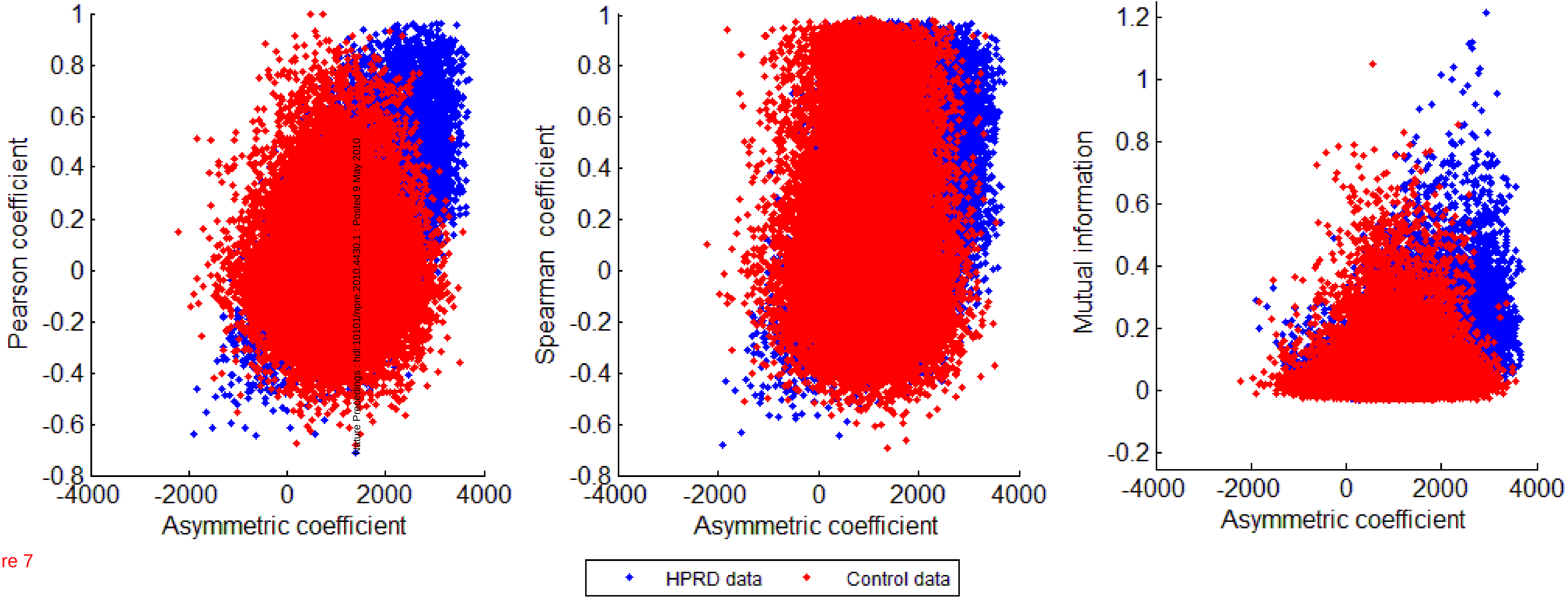

HPRD data
Number of pairs
10000
8000
6000
4000
2000
0
−2000
0
2000
4000
Asymmetric coefficient
Control data
Number of pairs
10000
8000
6000
4000
2000
0
−2000
0
2000
4000
Asymmetric coefficient

Figure 6



Pearson coefficient
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-4000
-2000
0
2000
4000
Asymmetric coefficient
Spearman coefficient
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-4000
-2000
0
2000
4000
Asymmetric coefficient
Mutual information
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-4000
-2000
0
2000
4000
Asymmetric coefficient
HPRD data
Control data

Figure 7

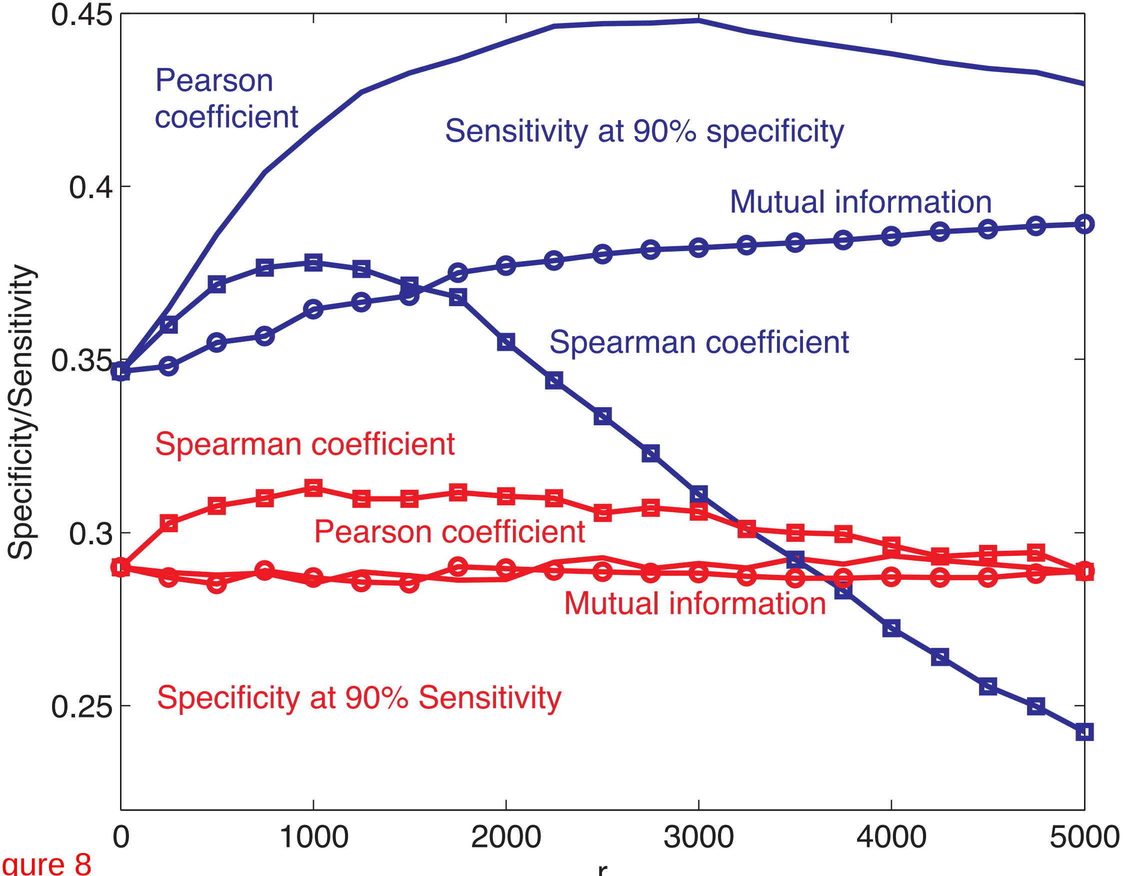

0.45
0.4
0.35
0.3
0.25
Specificity/Sensitivity
0
1000
2000
3000
4000
5000
r
Pearson coefficient
Sensitivity at 90% specificity
Mutual information
Spearman coefficient
Spearman coefficient
Pearson coefficient
Mutual information
Specificity at 90% Sensitivity

Figure 8