CHAPTER 4

Integration of multiple genome-wide datasets and analysis of functional relationships among genes in the human genome

IGOR B. KUZNETSOV[1] AND MOHAMMED J. ZAKI[2]

[1] *Gen*NY*sis Center for Excellence in Cancer Genomics,*
Department of Epidemiology and Biostatistics,
University at Albany, State University of New York,
One Discovery Drive, Rensselaer, NY 12144, USA
[2] *Department of Computer Science, Rensselaer Polytechnic Institute,*
Troy, NY 12180, USA

Genome studies are, by definition, very data intensive. Various high-throughput data acquisition techniques provide us with a multitude of large-scale datasets, such as whole genome sequences, gene expression, protein-protein interactions, functional pathways, location of transcription factor binding sites, etc. These datasets are not only large but also highly heterogeneous. Making biologically meaningful inferences requires a simultaneous computational analysis of these diverse genome-wide datasets. Such analyses, in turn, require linking heterogeneous data sets (data integration) and representing them in a form suitable for computational studies (data representation). These tasks are addressed by bioinformatics, a highly interdisciplinary field of science that, among its other goals, aims to integrate diverse and mutually complementary sources of 'omic' data (*e.g.*, genomic, proteomic, etc.) into a single coherent systems biology framework in order to provide functional inference, reveal essential features of gene and protein interaction networks, and ultimately to model these networks. The results of such integrative studies have several key advantages (Gerstein *et al*., 2002; Greenbaum *et al*., 2001; Vidal, 2001). In particular, using multiple sources of information may allow us to reduce systematic noise inherently present in all types of experimental data. Integrative approaches are also important for the studies of complex diseases, such as cancer (Wachi *et al*., 2005), since predicting the status of disease cases based on multiple biomarkers represents a starting point towards translating genomics research into clinical medicine. The integrative approach can also be used for predicting properties of one type of data based on other types of 'omic' (genomic, proteomic, etc) data (Greenbaum *et al*., 2001; Drawid and Gerstein, 2001; Qian *et al*., 2003; Zhang *et al*., 2004), for evaluating 'omic' datasets (Bader *et al*., 204), and for functional prediction and inference (Goh *et al*., 2006; Gunsalus *et al*., 2005; Lee *et al*., 2004).

Such a promise of the integrative approach is based on the general assumption that, within a given genome, there exist inter-relationships between heterogeneous types of genomic data (Grigoriev, 2001). Since even seemingly different data types describe various functional aspects of the same genome (*e.g.*, the human genome), it seems reasonable to anticipate the existence of non-random associations among them. However, the existence of such associations needs to be verified and their strength needs to be quantified for each particular combination of data types (Vidal, 2001). In this chapter, we use the term 'association' instead of 'correlation' in order not to confuse it with correlation between expression profiles.

A number of studies have demonstrated the existence of non-random pairwise associations between different types of large-scale 'omic' datasets. 'Non-random association' or just 'association' in this context means that genes that are functionally related with respect to one data type also tend to be related with respect to another data type. For the first time, such an association was demonstrated on the example of the yeast interactome and transcriptome. Since interacting proteins must be present within the cell at the same time, genes that encode them should also be expressed during the same time intervals. Consistent with this reasoning, it was shown that yeast genes with similar expression profiles are more likely to encode interacting proteins than randomly chosen genes (Ge *et al*., 2001). A related study of the yeast genome showed that genes encoding interacting proteins exhibit higher than average co-expression (Grigoriev, 2001). This study also showed that the yeast protein-protein interaction (PPI) dataset contains a larger proportion of strongly co-expressed proteins, compared to their baseline proportion in the entire yeast proteome. Similarly, yeast proteins from the same protein complex show a stronger co-expression than random proteins (Jansen *et al*., 2002). The interactome-transcriptome correlation demonstrated in yeast was also demonstrated for multicellular organism, *C. elegans* (Gunsalus *et al*., 2005; Walhout *et al*., 2002; Li *et al*., 2004). Another important type of association is that between expression and transcription factors (TF). It was shown for the yeast genome that when the same TFs target the same genes, these genes exhibit stronger co-expression than randomly selected ones (Yu *et al*., 2003).

The existence of two associations, PPI-expression and expression-TF locations, implies that there should also exist an association between PPI and TF locations. Consistent with this expectation, it was shown for proteins from the human *N*-methyl *D*-aspartate (NMDA) receptor that regulatory regions of the genes that encode interacting proteins are targeted by similar sets of TFs (Hannenhalli and Levy, 2003; Alter and Golub, 2004). The correlation between PPI and TF data was also employed in order to discover cooperative TF pairs that synergistically influence the expression of proteins that are located close to each other in the yeast protein-protein interaction network (Nagamine *et al*., 2005). Correlations that involve biological pathways were also studied. Since genes that belong to the same pathway are functionally related, they can be expected to be co-expressed and co-regulated. An association between pathways and expression was shown for both tumor (Yang *et al*., 2004; Huang and Wallqvist, 2006) and normal cells (Huang and Wallqvist, 2006) from the human genome. An association between pathway data and data on transcriptional regulation was also demonstrated for several selected human pathways (Hannenhalli and Levy, 2003). In yeast, relationships in a combination of three or more heterogeneous types of genome-wide datasets have also been studied (Tanay *et al*., 2004; Hwang *et al*., 2005; Carmona-Saez *et al*., 2006).

Most integrative studies have been done on the example of the yeast genome. Because of its relative simplicity, yeast is the best experimentally characterized eukaryotic organism for which many experimental large-scale datasets, such as PPI and locations of transcription factor binding sites (TFBS), are readily available. The human genome, on the other hand, is much more complex in nature and significantly harder to study experimentally. For instance, no comprehensive experimental datasets on protein-protein interactions and TFBS locations are yet available for the human genome. Due to the absence of such experimental datasets, information about multiple genome-wide associations that involve PPI and TFBS locations in the human genome is lacking. A possible way to overcome this limitation is to study associations using computationally inferred genome-wide datasets.

In this work, we use a novel computational approach to perform a comprehensive analysis of four types of data that describe the following functional features of the human genome: functional pathways, expression profiles, inferred protein-protein interactions, and inferred locations of transcription factor binding sites. We use inferred protein-protein interactions from OPHID (Online Predicted Human Interactome Database), the largest publicly available PPI database (Brown and Jurisica, 2005) that includes 8,687 human proteins. This PPI dataset is more than two orders of magnitude larger than the dataset of only 76 proteins used in a previously reported study of correlations involving the human interactome (Hannenhalli and Levy, 2003). We analyze types of associations that have not been studied previously for the human genome, including associations between expression and TFBS locations, PPI and expression, and pathway information and PPI. We study associations not only in pairwise combinations, but also in combinations of three and four data types.

Genome-wide datasets

This work deals with multiple heterogeneous sources of genomic data. We therefore need to use consistent unique gene identifiers for each of these sources. We utilize the human genome annotation version 38 from the Ensembl database (Hubbard *et al*., 2005) to assign a unique id to each gene and keep this id for each data type. We obtained the following four types of data for the human genome using publicly available sources (Table 1):

Table1. *Four genome-wide datasets used in this study.*

Data type	Description	Source	Genes
K	Functional pathways	KEGG (Kanehisa and Goto, 2000)	4,024
P	Protein-protein interactions	OPHID (Brown and Jurisica, 2005)	8,687
R	Expression profiles	SymAtlas (Su *et al*., 2005)	12,306
T	Putative TFBS found in the promoter regions	Ensembl (Hubbard *et al*., 2005), TRANSFAC (Matys *et al*., 2003)	23,326

1. Biological pathways from the KEGG database (Kanehisa and Goto, 2000). In KEGG, each gene from the human genome is assigned to one or more functional pathways. By mapping KEGG identifiers onto Ensembl identifiers, we generated a list of 4,024 genes for which pathway annotation is available.
2. Protein-protein interactions (PPI) from the OPHID database (Brown and Jurisica, 2005). OPHID catalogs human protein-protein interactions that are either determined experimentally or inferred from known protein-protein interactions in model organisms (*S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *M. musculus*). By mapping OPHID identifiers onto Ensembl identifiers, we generated a list of 8,687 genes whose protein products are annotated in the OPHID database.
3. Gene expression data from the SymAtlas database (Su *et al*., 2004). SymAtlas reports genome-scale gene expression measurements for 73 normal human tissues and 6 disease state tissues hybridized to Affymetrix HG-U133A array. Two replicates were used for each tissue. In our analysis, we excluded disease state tissues and only used 73 normal tissues. Expression levels for each tissue were averaged over the two replicates. Thus, each gene was represented by an expression profile that consists of 73 data points. By mapping Affymetrix identifiers onto Ensembl identifiers, we generated a list of 12,306 genes whose expression profiles are annotated in the SymAtlas database.
4. The data on transcription factor binding sites (TFBS) were obtained as follows. First, we used the Ensembl human genome assembly version 38 (Hubbard *et al*., 2005) to retrieve regulatory upstream region of each gene. We define regulatory upstream region as a 2KB region upstream of the transcription start site. In Ensembl, a gene can be annotated as producing multiple transcripts, 1.3 transcripts per gene on average (Curwen *et al*., 2004). In cases when more than one transcript is annotated for a given gene, we use known transcript with most 5' transcription start site. We choose known transcripts over novel transcripts because the former have more supporting evidence that the latter (Curwen *et al*., 2004). We used this procedure to retrieve regulatory upstream regions of all protein-coding genes (a total of 23,326 genes). Second, we used the Match software program (Kel *et al*., 2003) to scan the upstream regions for TFBS annotated in the TRANSFAC database (Matys *et al*., 2003). The TRANSFAC database is a library of experimentally identified transcription factor binding sites represented in the form of a position weight matrix (PWM). Match is a tool that searches for putative TFBS in input DNA sequences by using a library of PWMs. Match was run using the library of high-quality vertebrate PWMs and the option to minimize the number of false positives. By parsing Match output, we obtained a list of putative TFBS found in the upstream regions of 23,326 human genes.

Conversion of the datasets into a unified matrix format

Each type of genomic data was converted into a unified matrix format. In this format, a symmetric *n* by *n* matrix numerically summarizes a particular type of functional relationships observed among *n* genes. Each of the four types of data described above was converted into a matrix format as follows:

1. KEGG pathways are represented by matrix K (size 4,024 × 4,024). An element k_{ij} in K matrix is equal to 1 if products of genes *i* and *j* belong to at least one common KEGG pathway and 0 otherwise.
2. Protein-protein interactions are represented by matrix P (size 8,687 × 8,687). An element p_{ij} in P matrix has a binary value of 1 or 0, indicating the presence or absence of protein-protein interaction between products of genes *i* and *j*.

3. Expression profiles are represented by matrix R (size 12,306 × 12,306). An element r_{ij} in R matrix is the Pearson correlation coefficient (PCC) between expression profiles of genes *i* and *j*. For the cases when at least one gene in a pair (*i,j*) is mapped onto multiple Affymetrix probe sets (3,837 out of 12,306 genes), we calculate PCC between all probe set pairs that correspond to (*i,j*) and choose a PCC with the largest magnitude. Negative correlations in R matrix were set to zero. For analyses that involve computing association scores (see below), we use a binary version of R matrix in which all elements that have values equal to or greater than 0.7 (strong correlation) are set to 1 and all elements that have values below 0.7 are set to 0.
4. The *cis*-similarity between promoter regions of genes is represented by matrix T (size 23,326 × 23,326). An element t_{ij} in T matrix is the number of unique TFBSs observed in the promoter regions of both gene *i* and *j*. Unique means that all occurrences of binding sites for the same TF are counted only once for each promoter region. For instance, if the promoter region of gene *i* contains 4 sites for transcription factor A and 1 site for transcription factor B, whereas the promoter region of gene *j* contains 2 sites for transcription factor A and 3 sites for transcription factor B, the value of t_{ij} will be equal to 2. The idea of this definition of *cis*-similarity is to attempt to account for the number of common transcription factors that control both gene *i* and *j*.

When we study a combination of two or more types of data, we only use genes for which all types of required annotation are available and exclude genes with missing annotation. For example, when we study associations between K and P matrices, we take a set of genes for which both KEGG pathway and protein-protein interaction data are annotated.

Statistical significance of associations among multiple data types

There are two main ideas behind presenting a particular type of genomic data as a symmetric matrix that describes a certain type of functional relationship between gene pairs. One idea is to reveal statistically significant functional associations among multiple matrices by using multiplication of equivalent matrix elements. The other idea is to use a matrix to construct a graph that displays the strength of relationships among genes. In such a graph, a pair of functionally related genes is represented by two connected nodes corresponding to a non-zero matrix element (see below). In general, when elements from *k* matrices of dimension *n*, $M_1 \ldots M_k$, that repr esent *k* types of genomic data for *n* genes are multiplied, and a final matrix is obtained, $F[i,j]=M_1[i,j]*\ldots*M_k[i,j]$ (note that this is an element-wise multiplication, not a conventional matrix product). In this final matrix *F*, gene pairs that exhibit strong associations across all *k* types of data will correspond to elements with large absolute value. The overall strength of functional associations within a group of *n* genes represented by *k* matrices can be quantified by computing the sum of all elements in the final matrix, *S(n,k),* as follows:

$$S(n,k) = \sum_{i<j}^{n} M_1[i,j] \cdot M_2[i,j] \cdot \ldots \cdot M_k[i,j] \tag{1}$$

If *S(n,k)* is significantly higher than that expected by chance, it will indicate that genes in the multiplied matrices exhibit a strong non-random association across *k* types of genomic data. We estimate the statistical significance of *S(n,k)* by comparing it to the distribution of random scores. A random score is obtained by randomly permuting elements in each matrix $M_1,\ldots,M_k$ and then using these permuted matrices to obtain a score according to Eq.1. For each matrix combination we generate 10,000 random scores. The p-value of the observed score, $P(R(n,k)\geq S(n,k))$, is computed as follows:

$$P(R(n,k) \geq S(n,k)) = \frac{N(R(n,k) \geq S(n,k))}{10{,}000} \tag{2}$$

where *R(n,k)* is random score and $N(R(n,k)\geq S(n,k))$ is the number of random scores that are equal to or larger than *S(n,k).* We applied the Shapiro-Wilk normality test and found random association scores to be normally distributed (data not shown). Histograms of the distributions of random scores can be found in Supplementary information. Since most p-values obtained from random simulations are zero, we use the z-score to rank the associations:

$$z - score(S(n,k)) = \frac{S(n,k) - < R(n,k) >}{\sigma_{R(n,k)}} \tag{3}$$

where *<R(n,k)>* is the average and $\sigma_{R(n,k)}$ is the standard deviation of the random score.

Functional associations observed in multiple types of genome-wide datasets

In this section, we perform a qualitative study of genome-wide associations observed among the four types of genomic data. The idea of this study is to examine whether the properties of genes with respect to one type of functional data are correlated with other types of functional data. For example, we can classify pairs of genes into interacting and non-interacting categories and examine the average correlation coefficient between their expression profiles in order to see whether expression profiles of genes whose products interact tend to have a higher correlation coefficient than the profiles of non-interacting ones. Here, we study global genome-wide associations for the following combinations of data types: K-P, R-T, K-R, P-T, K-P-R, and K-P-T and demonstrate the existence of potentially significant relationships observed among these data types. A rigorous statistical analysis of the significance of associations for all possible combinations of data types that confirms the qualitative trends discussed here is presented in the following sections of the manuscript.

First, we analyze the associations between functional pathways and protein-protein interactions (K-P association). The comparison of pathway information for interacting and non-interacting proteins shows that 40.82% of interacting protein pairs share at least one functional pathway (meaning that both proteins in the pair belong to the same pathway), whereas only 5.55% of non-interacting protein pairs share pathway annotation (Figure 1A). This means that interacting protein pairs are seven times more likely to participate in the same pathway than non-interacting protein pairs. Analysis of the reverse relation shows that if two proteins participate in the same pathway, they are eleven times more likely to interact than proteins from different pathways (Figure 1B).

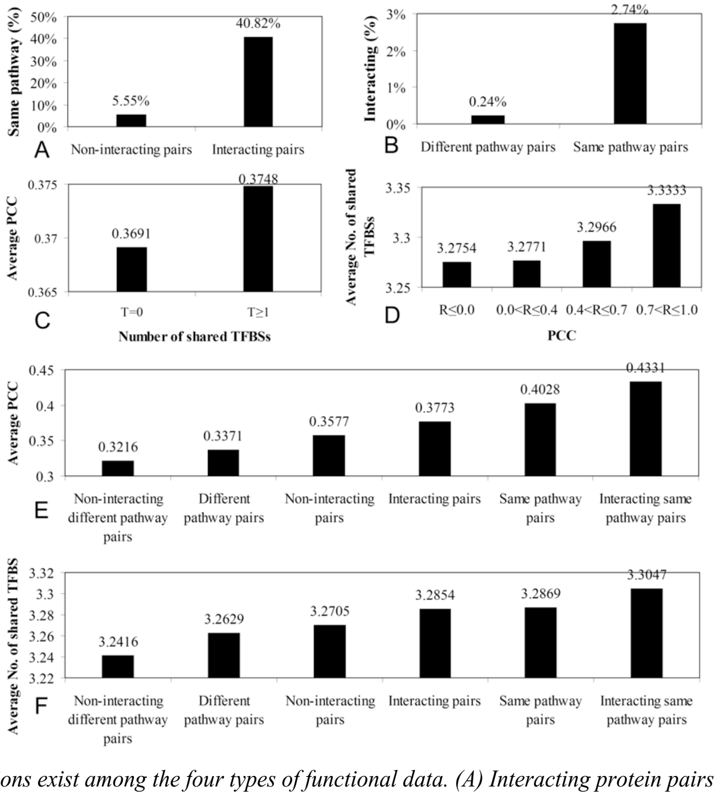

Figure1. *Associations exist among the four types of functional data. (A) Interacting protein pairs are more likely to participate in the same pathway than non-interacting protein pairs. (B) Protein pairs from the same pathway are more likely to interact than protein pairs from different pathways. (C) A pair of genes that share common TFBS in the promoter regions shows a higher correlation between expression profiles than a pair without any shared TFBS. (D) An increase in correlation between expression profiles is associated with an increase in the number of shared TFBS. (E) Pairs of proteins from the same pathway and/or pairs of interacting proteins are more likely to show correlated expression. (F) Pairs of protein from the same pathway and/or pairs of interacting proteins are more likely to share common TFBS in their promoter regions.*

Second, we analyze the associations between co-expression and *cis*-similarity of promoter regions (R-T association). This analysis shows that, on average, correlation between expression profiles of genes that share common TFBS is higher (PCC=0.375) than that between expression profiles of genes that do not share any common TFBS (PCC=0.369) (Figure 1C). Analysis of the reverse relation shows that an increase in the level of co-expression of gene pairs is associated with an increase in the number of common TFBS found in their promoter regions (Figure 1D). These results confirm to an empirical expectation that co-expressed genes should have similar *cis*-profiles and *vice versa*. However, the trends shown in Figure 1C and 1D are very subtle and their statistical significance is not obvious. One possible reason of weak trends is that the computational identification of putative TFBS via sequence motif-based methods is inherently prone to noise because of a very high percentage of false positive predictions (Robinson *et al.*, 2006).

Third, we analyze the following three types of associations: K-R, P-R, and K-P-R. We divided all gene pairs into six categories according to whether their products are interacting and/or participating in same functional pathways and compared the average correlations between expression profiles for these six categories (Figure 1E). From right to left in Figure 1E, the largest average PCC between expression profiles is found for gene pairs that both interact and participate in same pathways (PCC = 0.4331), whereas the smallest average PCC is found for gene pairs that neither interact nor participate in same pathways (PCC = 0.3216). We also observe that the average PCC is higher for gene pairs that participate in same pathways (PCC = 0.4028) than for interacting pairs (PCC = 0.3773). These observations suggest that, with respect to concerted expression, genes from the same pathway act as a more cohesive biological module than genes producing physically interacting proteins. Experimental evidence shows that interacting proteins from the same complex are not necessarily produced by co-regulated genes. For example, cyclin-dependent kinase and cyclin together form a protein complex. While the former is produced from a constantly transcribed gene, the latter is produced in a regulated manner (Ge *et al.*, 2001).

Fourth, we analyze K-T, P-T, and K-P-T associations by computing the average number of common TFBS for the same six categories of gene pairs described above. The results of this analysis, shown in Figure 1F, reveal a trend very similar to the one shown in Figure 1E: the largest number of common TFBS is observed for gene pairs that both interact and participate in the same pathways, whereas the smallest number of common TFBS is observed for gene pairs that neither interact nor participate in the same pathway. These two related trends indirectly indicate that the level of co-expression (measured by PCC) and the *cis*-similarity (measured by the number of common TFBS) are correlated with each other, which is in agreement with the direct relationship between them shown in Figure 1C and 1D. The small differences in the number of common TFBS observed in Figure 1F can be attributed to the fact that the computational procedure for the identification of putative TFBS produces a very large number of false positives (Robinson *et al.*, 2006).

Statistical significance of functional associations observed in multiple datasets

The qualitative analyses shown in the previous section indicate the existence of potentially significant associations among various types of 'omic' data. In this section, we use a rigorous quantitative approach to evaluate the statistical significance of the observed associations on the genome-wide scale. Given two or more matrices that represent particular types of 'omic' data for a group of genes, we measure the strength of association among these data types by means of an association score. This association score is defined as the sum of products between corresponding elements of the matrices under consideration (Eq. 1). Statistical significance of the observed association score is estimated by comparing it to the distribution of random association scores obtained from randomly permuted matrices. Since we have four matrices that correspond to the four types of data, there are eleven possible combinations of two, three and four matrices. P-values and z-scores for each combination are shown in Table 2. Histograms of all random distributions can be found in Figure 2.

Table2. *P-values and z-scores estimated from the random permutation experiment.*

Data analyzed	Type of association	Genes	p-value	z-score
Pairwise combinations	K-P	2,424	0	160.97
	K-R	3,271	0	85.35
	R-T	11,775	0	75.63
	P-R	6,784	0	7.70
	K-T	3,939	0	7.59
	P-T	8,513	0.0498	1.66
Triple combinations	K-P-T	2,424	0	142.26
	K-P-R	2,154	0	82.22
	K-R-T	3,271	0	73.90
	P-R-T	6,784	0	7.47
Quadruple combination	K-P-R-T	2,154	0	75.65

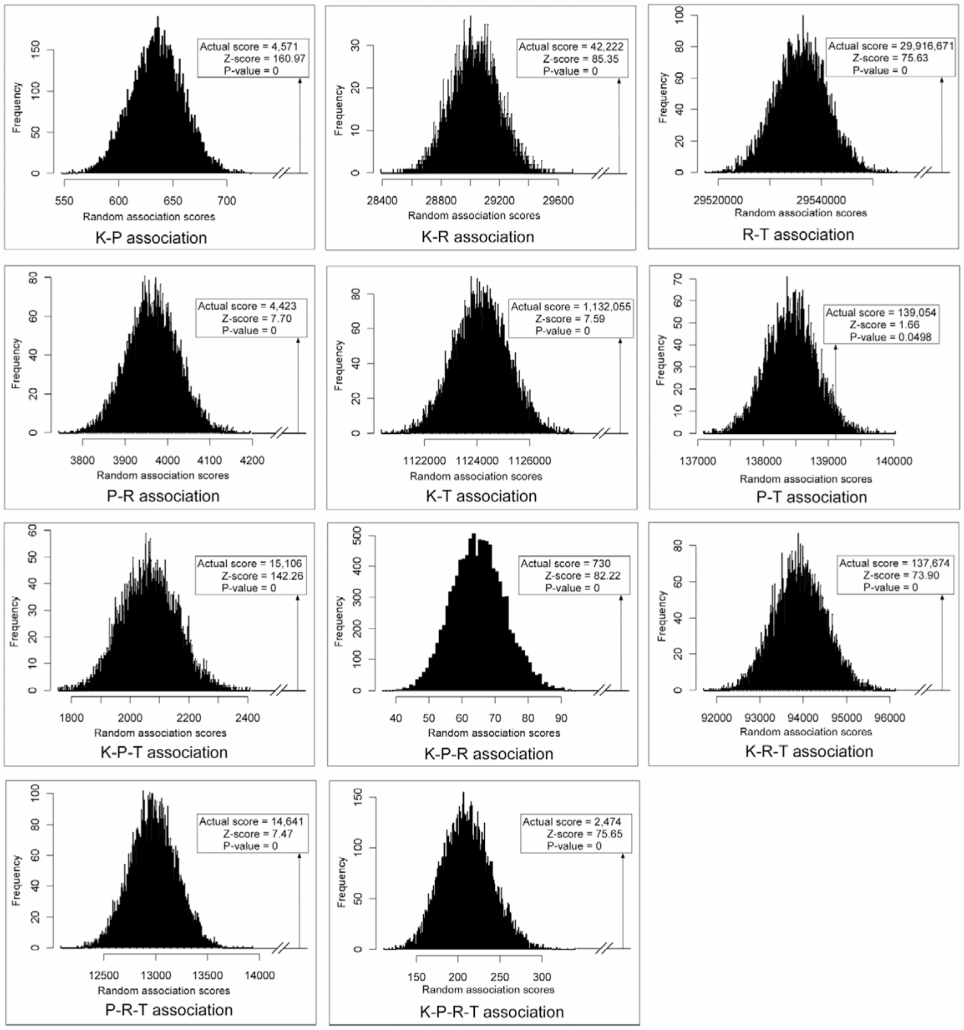

Figure2. *Distributions of random association scores*

The results shown in Table 2 indicate that all eleven combinations of data types demonstrate significant associations as evidenced by low *p*-values. Below, we briefly discuss biological implications of each association. The results for pairwise combinations indicate:

K-P (z-score=160.97, p=0) - the existence of a highly significant association between protein-protein interactions and protein function. Since in our methodology associations are not directional, K-P association is equivalent to P-K association, thus implying that interacting proteins tend to participate in the same functional pathway, and *vice versa*, proteins from the same functional pathway tend to interact.

K-R (z-score=85.35, p=0) - the existence of a highly significant association between co-expression and gene function. It shows that genes participating in the same pathway tend to be co-expressed, and *vice versa*, co-expressed genes tend to participate in the same pathway.

R-T (z-score=75.63, p=0) - co-expressed genes tend to share similar *cis*-profiles, and *vice versa*, genes with similar *cis*-profiles tend to be co-expressed.

P-R (z-score=7.70, p=0) - genes that encode interacting proteins tend to be co-expressed, and *vice versa*, co-expressed genes tend to encode interacting proteins.

K-T (z-score=7.59, p=0) - genes from the same pathway tend to have similar *cis*-profiles, and *vice versa*, genes with similar *cis*-profiles tend to participate in the same pathway.

P-T (z-score=1.66, p=0.0498) - the existence of a marginally significant association between protein-protein interactions and the similarity of *cis*-profiles of the genes that encode interacting proteins.

The results for all combinations of three data types, described below, also demonstrate highly statistically significant genome-wide associations:

K-P-T (z-score=142.26, p=0) - the existence of a highly significant association that links gene function, K, interactions between gene products, P, and *cis*-similarity of the promoter regions, T. This association implies that genes from the same pathway both tend to code for interacting proteins and share a similar set of TFs in their promoter regions.

K-P-R (z-score=82.22, p=0) - the existence of a highly significant association that links gene function, K, interactions between gene products, P, and co-expression, R. Biologically, this association is similar to K-P-T association and implies that genes from the same pathway both tend to code for interacting proteins and to be co-expressed.

K-R-T (z-score=73.90, p=0) - the existence of a highly significant association that links gene function, K, co-expression, R, and *cis*-similarity of the promoter regions, T. This association implies that genes from the same pathway tend to be both co-expressed and have a similar set of TFs in their promoter regions.

P-R-T (z-score=7.47, p=0) - the existence of a significant association that links interactions between gene products, P, co-expression, R, and *cis*-similarity of the promoter regions, T. This association implies that genes whose products interact tend to be co-expressed and have a similar set of TFs in their promoter regions. However, it should be pointed out that the strength of P-R-T association is much weaker than that of other triple associations as indicated by a considerably lower z-score.

Finally, the results for the combination of all four data types, K-P-R-T, indicate that this quadruple association is also highly significant (z-score=75.65, p=0). This association indicates that genes from the same pathway simultaneously tend to encode interacting proteins, be co-expressed, and have a similar set of TFs in their promoter regions.

Pathway-level analysis of functional associations

The analysis reported in the previous section summarizes global genome-wide relations among data types by considering all genes in the human genome simultaneously. A similar analysis can be performed by considering a group of genes that belong to a particular functional category. A good example of a functional category is a functional pathway, which can be considered as a biological module that carries out a specific genomic function. Depending on the function of the pathway, one may expect certain pathway-specific associations to be more pronounced than the others. In this section, we use the classification of functional pathways from the KEGG database. The main difference from the global analysis reported in the previous section is that pathway-level analysis is done for a group of functionally related genes that belong to a particular KEGG pathway. This kind of analysis enables us to categorize the associations between the pathway data and other types of genomic data and to determine which types of associations are most profound in a particular functional category. Since the TFBS data seem to be noisy and therefore least reliable, we use only the PPI data and the gene expression data for the pathway-level analysis. This leaves us with three combinations to analyze for pathway-level associations: K-P, K-R and K-P-R. These three combinations provide the following biological information for a given pathway: K-P describes its relative enrichment in interacting proteins, K-R describes its relative enrichment in co-expressed genes, and K-P-R describes its relative enrichment in genes that are both co-expressed and code for interacting proteins. When we use Eqs.2-3 to analyze a group of *m* genes that belong to a particular pathway, each random *m* by *m* matrix for a given type of data is obtained by randomly sampling, without replacement, *m* genes from a list of all human genes annotated with this particular data. Pathways containing less than five annotated genes were excluded from this analysis.

Out of 174 pathways annotated in the KEGG database for the human genome, we identified 98 pathways that are significantly (p<0.05) enriched in interacting proteins, 34 pathways that are significantly enriched in co-expressed genes, and 75 pathways that are significantly enriched in genes that are both co-expressed and code for interacting proteins. Lists for all pathways and all combinations, ranked by z-score, are given in Table 3, 4 and 5. It should be noted that when pathways are analyzed with respect to concomitant enrichment in co-expressed genes whose protein products also interact (the triple K-P-R association), several additional pathways emerge as significant (Table 5). For example, two pathways ('Glutamate metabolism' and 'Glutathione metabolism') are identified as showing significant concomitant enrichment, even though they do not show enrichment in interacting proteins or co-expressed genes. Some pathways are concomitantly enriched even though they show enrichment in either interacting proteins or co-expressed genes, but not both. For example, 'Cholera' pathway shows very strong concomitant enrichment (top 5th in Table 5), but it does not show enrichment in co-expressed genes. Similarly, 'Olfactory transduction' pathway shows a significant concomitant enrichment without being enriched in interacting proteins. These observations indicate that combining multiple types of genomic data reveals additional functional features of individual pathways that cannot be revealed by studying simple pairwise associations.

Table 3. Pathways enriched in interacting proteins.

KEGG ID	z-score	p-value	Pathway name	KEGG category
hsa03050	237.19	0	Proteasome	Genetic Information Processing
hsa03020	134.22	0	RNA polymerase	Genetic Information Processing
hsa03010	86.94	0	Ribosome	Genetic Information Processing
hsa00193	61.66	0	ATP synthesis	Metabolism
hsa00240	48.63	0	Pyrimidine metabolism	Metabolism
hsa04110	44.15	0	Cell cycle	Cellular Processes
hsa03022	40.98	0	Basal transcription factors	Genetic Information Processing
hsa04130	39.30	0	SNARE interactions in vesicular transport	Genetic Information Processing
hsa00020	39.19	0	Citrate cycle (TCA cycle)	Metabolism
hsa04350	34.76	0	TGF-beta signaling pathway	Environmental Information Processing
hsa00970	33.41	0	Aminoacyl-tRNA biosynthesis	Genetic Information Processing
hsa00230	33.08	0	Purine metabolism	Metabolism
hsa04660	30.70	0	T cell receptor signaling pathway	Cellular Processes
hsa04664	29.88	0	Fc epsilon RI signaling pathway	Cellular Processes
hsa04210	29.36	0	Apoptosis	Cellular Processes
hsa05010	28.28	0	Alzheimer's disease	Human Diseases
hsa04662	26.97	0	B cell receptor signaling pathway	Cellular Processes
hsa04650	26.85	0	Natural killer cell mediated cytotoxicity	Cellular Processes
hsa05040	26.80	0	Huntington's disease	Human Diseases
hsa04510	25.42	0	Focal adhesion	Cellular Processes
hsa04630	24.95	0	Jak-STAT signaling pathway	Environmental Information Processing
hsa04610	23.97	0	Complement and coagulation cascades	Cellular Processes
hsa00190	23.76	0	Oxidative phosphorylation	Metabolism
hsa04920	23.34	0	Adipocytokine signaling	Cellular Processes

			pathway	
hsa05120	23.30	0	Epithelial cell signaling in Helicobacter pylori infection	Human Diseases
hsa05110	23.02	0	Cholera	Human Diseases
hsa00252	22.89	0	Alanine and aspartate metabolism	Metabolism
hsa04512	22.13	0	ECM-receptor interaction	Environmental Information Processing
hsa04520	22.09	0	Adherens junction	Cellular Processes
hsa04010	21.83	0	MAPK signaling pathway	Environmental Information Processing
hsa04330	21.63	0	Notch signaling pathway	Environmental Information Processing
hsa04310	20.44	0	Wnt signaling pathway	Environmental Information Processing
hsa05020	20.21	0	Parkinson's disease	Human Diseases
hsa00220	20.18	0	Urea cycle and metabolism of amino groups	Metabolism
hsa03030	20.14	0	DNA polymerase	Genetic Information Processing
hsa05030	19.83	0	Amyotrophic lateral sclerosis (ALS)	Human Diseases
hsa04320	19.55	0	Dorso-ventral axis formation	Cellular Processes
hsa04910	19.48	0	Insulin signaling pathway	Cellular Processes
hsa04810	19.48	0	Regulation of actin cytoskeleton	Cellular Processes
hsa04620	19.36	0	Toll-like receptor signaling pathway	Cellular Processes
hsa04710	17.90	0	Circadian rhythm	Cellular Processes
hsa00620	17.69	0	Pyruvate metabolism	Metabolism
hsa04120	17.26	0	Ubiquitin mediated proteolysis	Genetic Information Processing
hsa04670	17.19	0	Leukocyte transendothelial migration	Cellular Processes
hsa04930	16.28	0	Type II diabetes mellitus	Human Diseases
hsa04370	15.93	0	VEGF signaling pathway	Environmental Information Processing
hsa05050	15.09	0	Dentatorubropallidoluysian atrophy (DRPLA)	Human Diseases
hsa03060	14.11	0.0001	Protein export	Genetic Information Processing
hsa00563	13.96	0	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	Metabolism
hsa04540	13.80	0	Gap junction	Cellular Processes
hsa04360	13.73	0	Axon guidance	Cellular Processes
hsa04150	13.16	0	mTOR signaling pathway	Environmental Information Processing
hsa00720	13.15	0.0001	Reductive carboxylate cycle (CO2 fixation)	Metabolism
hsa00100	11.62	0.0001	Biosynthesis of steroids	Metabolism
hsa00010	11.21	0	Glycolysis / Gluconeogenesis	Metabolism
hsa04730	11.12	0	Long-term depression	Cellular Processes
hsa05130	10.44	0	Pathogenic Escherichia coli infection	Human Diseases
hsa00130	9.95	0.0015	Ubiquinone biosynthesis	Metabolism
hsa00290	9.85	0.0007	Valine, leucine and isoleucine biosynthesis	Metabolism
hsa00860	9.82	0.0002	Porphyrin and chlorophyll metabolism	Metabolism
hsa00271	9.77	0	Methionine metabolism	Metabolism
hsa04060	9.74	0	Cytokine-cytokine receptor interaction	Environmental Information Processing
hsa00500	9.72	0	Starch and sucrose	Metabolism

			metabolism	
hsa04514	9.45	0	Cell adhesion molecules (CAMs)	Environmental Information rocessing
hsa00920	9.44	0.0111	Sulfur metabolism	Metabolism
hsa04612	9.37	0	Antigen processing and presentation	Cellular Processes
hsa00790	9.14	0.0002	Folate biosynthesis	Metabolism
hsa00330	9.13	0	Arginine and proline metabolism	Metabolism
hsa04340	9.01	0	Hedgehog signaling pathway	Environmental Information Processing
hsa04720	8.97	0	Long-term potentiation	Cellular Processes
hsa04530	8.91	0	Tight junction	Cellular Processes
hsa00640	8.84	0.0001	Propanoate metabolism	Metabolism
hsa04640	8.65	0	Hematopoietic cell lineage	Cellular Processes
hsa05060	7.89	0.0008	Prion disease	Human Diseases
hsa00400	7.63	0.0012	Phenylalanine, tyrosine and tryptophan biosynthesis	Metabolism
hsa00710	7.51	0.0002	Carbon fixation	Metabolism
hsa00650	7.05	0.0002	Butanoate metabolism	Metabolism
hsa00030	6.93	0.0005	Pentose phosphate pathway	Metabolism
hsa04950	6.63	0.0011	Maturity onset diabetes of the young	Human Diseases
hsa04020	6.57	0	Calcium signaling pathway	Environmental Information Processing
hsa00670	6.49	0.0023	One carbon pool by folate	Metabolism
hsa00630	6.47	0.0055	Glyoxylate and dicarboxylate metabolism	Metabolism
hsa00071	5.26	0.0016	Fatty acid metabolism	Metabolism
hsa00260	5.25	0.0016	Glycine, serine and threonine metabolism	Metabolism
hsa04742	5.21	0.0037	Taste transduction	Cellular Processes
hsa04140	5.21	0.0026	Regulation of autophagy	Genetic Information Processing
hsa04940	4.98	0.0027	Type I diabetes mellitus	Human Diseases
hsa00040	4.84	0.0370	Pentose and glucuronate interconversions	Metabolism
hsa02010	4.77	0.0052	ABC transporters	Environmental Information Processing
hsa00564	4.56	0.0027	Glycerophospholipid metabolism	Metabolism
hsa04080	4.34	0.0006	Neuroactive ligand-receptor interaction	Environmental Information Processing
hsa00280	4.22	0.0058	Valine, leucine and isoleucine degradation	Metabolism
hsa00150	4.13	0.0125	Androgen and estrogen metabolism	Metabolism
hsa00052	4.12	0.0077	Galactose metabolism	Metabolism
hsa00930	3.82	0.0222	Caprolactam degradation	Metabolism
hsa00450	3.77	0.0159	Selenoamino acid metabolism	Metabolism
hsa00510	3.15	0.0268	N-Glycan biosynthesis	Metabolism
hsa04070	2.71	0.0214	Phosphatidylinositol signaling system	Environmental Information Processing

Graph-theoretical analysis of gene networks

The analyses presented in the previous sections dealt with the overall statistical significance of multiple associations among genes from a given group (such as a specific pathway, for instance). However, characterizing genes involved in particular cellular processes requires not only an analysis of the overall strength of functional associations among these genes, but also an identification of the fine structure of the process-specific gene network(s) (Myers *et al.*, 2005). Reconstructing and modeling gene networks is one of the most challenging problems of genomic research. Usually, a gene network is described as a graph. A graph consists of a set of points, called nodes, along with a set of lines, called edges, which connect the nodes. Each edge connects two nodes. A sub-graph S of a graph V is a graph whose nodes and edges are also in V. A graph is said to be connected if there exists a path between any pair of nodes (Figure 3). In a graph describing a gene network each node represents a gene and the presence of an edge connecting two nodes indicates the existence of a functional association between the corresponding connected genes. An edge can mean the presence of either a direct physical interaction or an indirect functional association between gene products. The representation of genome-wide data in a unified matrix format described above is perfectly suited for the reconstruction of gene networks using a graph-theoretical approach. In this approach, a matrix element *M[i,j]* describes the strength of connection (association) between genes *i* and *j*. If the value of *M[i,j]* is 0, it means there is no edge (no connection) between genes *i* and *j*. For the sake of simplicity, we used a binary version of the co-expression matrix R, in which all elements that have value of 0.8 or higher were set to 1 (high co-expression) and all other elements were set to 0 (low co-expression). Matrices K and P are binary by definition. After multiplying these matrices we obtain a final binary matrix (F=P*R or F=P*R*K), which is subsequently supplied to a graph-mining algorithm that finds all connected sub-graphs in this final matrix. Graphs were visualized using the GOlorize plugin (Garcia *et al.*, 2007).

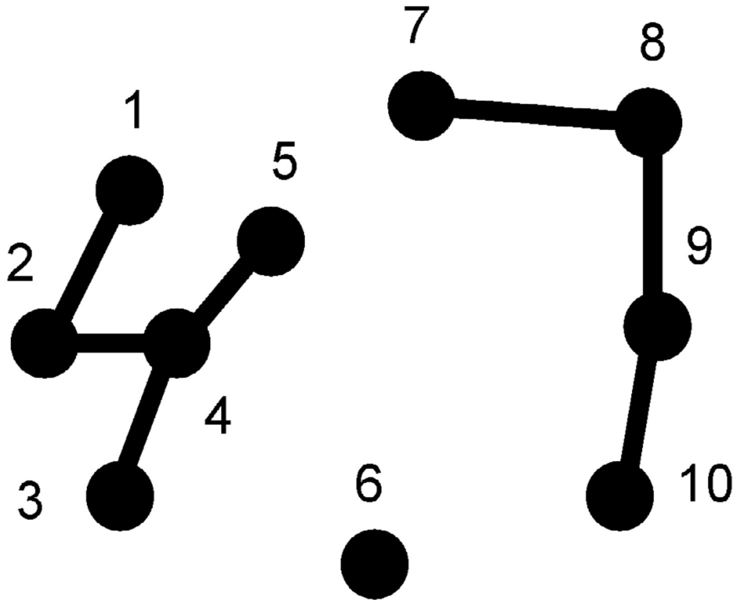

Figure3. *An example of a graph that consists of 10 nodes (numbered 1 through 10). There are three connected sub-graphs: 1st consists of nodes 1 through 5, 2nd consists of node 6, 3rd consists of nodes 7 through 10*

The largest connected sub-graph obtained from the analysis of the final matrix P*R is shown in Figure 4. It consists of 249 nodes (genes). In this graph, according to the definition of P*R multiplication, two genes are connected by an edge if they both are co-expressed and code for interacting proteins. There are 9 pathways significantly over-represented among these genes. These pathways include ATP synthesis, Epithelial cell signaling in *H. pylori* infection, Insulin signaling, MAPK signaling, Oxidative phosphorylation, Cholera, Toll-like receptor signaling, T-cell receptor signaling, and Adherense junction. Notably, most of them are signaling pathways. Analysis of the structure of this sub-graph shows that genes from most pathways (such as Adherense junction, MAPK signaling, Toll-like signaling, Epithelial cell signaling, Insulin signaling, and ATP synthesis) form a densely inter-connected network core. On the other hand, genes from Oxidative phosphorylation pathway, which is the largest gene group, form a loosely connected set of peripheral nodes all around the core. This example of network topology illustrates that, when a connection (edge) in a gene network is defined as a concerted expression of interacting proteins, a core set of nodes (proteins) involved in cell signaling is revealed. These core nodes transmit signals in concerted manner to numerous peripheral proteins involved in phosphorylation, which subsequently modify other proteins regardless of mutual co-expression. A very different picture is observed when the largest connected sub-graph from the final matrix P*R*K is analyzed (Figure 5). In this graph, according to the definition of P*R*K multiplication, two genes are connected by an edge if they are co-expressed and code for interacting proteins and belong to the same pathway. This sub-graph is smaller (56 nodes) and does not have many homogenous peripheral nodes. The number of over-represented pathways is larger (13) and they are more diverse than in the case of the P*R sub-graph, many being involved in extra-cellular interactions and immune response.

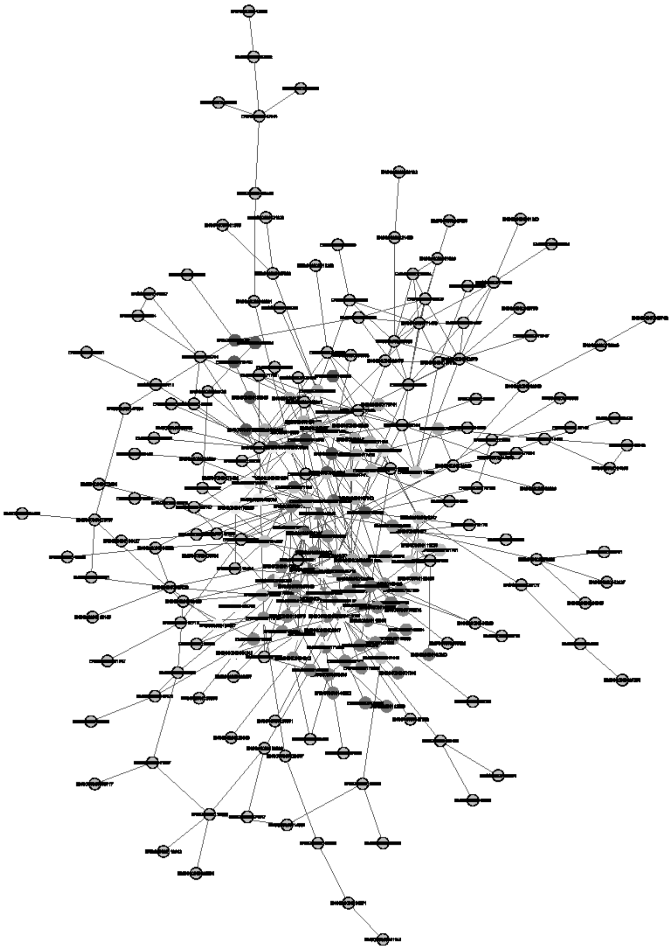

Figure 4. *The largest connected sub-graph found in the final P*R matrix (249 nodes). Each node is colored according to over-represented KEGG pathway(s) it belongs to. Over-represented means that the probability of observing the same or greater number of genes from a given pathway in a graph of the same size is less than 0.05. There are 9 over-represented pathways: ATP synthesis, Epithelial cell signaling in* H. pylori *infection, Insulin signaling, MAPK signaling, Oxidative phosphorylation, Cholera, Toll-like receptor signaling, T-cell receptor signaling, and Adherense junction*

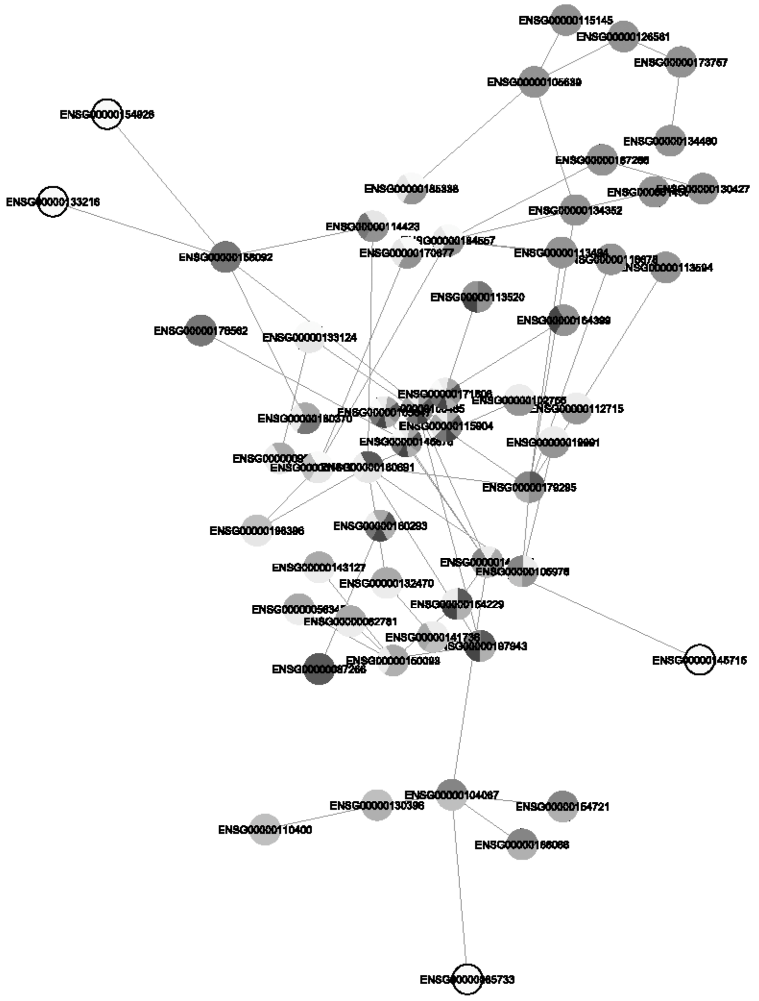

*Figure5. The largest connected sub-graph found in the final matrix P*R*K (56 nodes). Each node is colored according to over-represented KEGG pathway(s) it belongs to. Over-represented means that the probability of observing the same or greater number of genes from a particular pathway in a graph of the same size is less than 0.05. There are 13 over-represented pathways: Jak-STAT signaling, Focal adhesion, Insulin signaling, Fc epsilon RI signaling, Natural killer cell mediated cytotoxicity, Type II diabetes mellitus, T-cell receptor signaling, Leukocyte transendothelial migration, Regulation of actin cytoskeleton, Cytokine-cytokine receptor interaction, Epithelial cell signaling in* H. pylori *infection, Adherense junction, Dorso-ventral axis formation*

Discussion and Conclusion

In general, the results of the quantitative analysis of the genome-wide pairwise associations are consistent with the qualitative study performed on the same datasets (see Figure 1) and discussed in the first section of Results, thus confirming the utility of the proposed approach. For instance, Figure 1E shows that the average correlation between expression profiles is larger for gene pairs from the same pathway, K, (PCC=0.4028) than for gene pairs that encode interacting proteins (PCC=0.3773). This observation is consistent with a larger z-score observed for K-R association (z-score=85.35) compared to that for P-R association (z-score=7.70). Similarly, Figure 1F shows that the average number of shared TFBS is larger for gene pairs from the same pathway (3.2869) than for gene pairs whose products interact (3.2705). This observation is also consistent with a larger z-score observed for K-T association (z-score=7.59) compared to that for P-T association (z-score=1.66). If we assume that K and P matrices contain similar amounts of noise, then the observation that the z-scores for K-R and K-T associations are larger than those for P-R and P-T suggests that transcriptional co-regulation is more important for genes from the same pathway than for genes that encode interacting proteins. It should also be noted that, to the best of our knowledge, out of the six pairwise combinations of data types utilized in this work, three (P-R, R-T, and K-P) have never been studied for the human genome.

The strongest pairwise associations, indicated by very high z-scores, are observed for combinations involving pathway data, K-P and K-R. This observation is consistent with empirical expectations and confirms that genes from the same functional pathway tend to be co-expressed and code for interacting proteins. The only marginally significant genome-wide association is observed between PPI data and *cis*-similarity of promoter regions (P-T combination, p=0.0498, Table 2). The relatively low z-scores for two associations involving the T matrix (P-T and K-T) are not straightforward to interpret. On one hand, a large amount of noise present in matrix T may dampen real biological associations. On the other hand, the R-T association is quite significant (z-score=75.63) despite the noise present in the T matrix. The observation that P-R association, which is related to P-T, also has a relatively low z-score of 7.7 provides an additional argument in favor of the assumption that the weakness of the genome-wide P-T association may reflect a real biological phenomenon.

The application of our methodology to study associations in groups of genes from individual functional pathways shows that pathways enriched in interacting proteins (K-P association, Table 3) are mostly the ones for genetic information processing. These pathways tend to contain large protein complexes, such as the ribosome and DNA/RNA polymerases. Pathways enriched in co-expressed genes (K-R association, Table 4) are mostly the pathways for environmental information processing. These pathways can be thought of as biological modules whose genes need to be expressed in a concerted manner in response to external stimuli. Metabolic pathways seem to be under-represented in the list of pathways enriched in co-expressed genes. There are 112 metabolic pathways, comprising 64% of all 174 annotated pathways. However, out of the total of 34 pathways significantly enriched in co-expressed genes, only six (18%) are metabolic pathways. This observation is consistent with previously reported results that metabolic pathways do not show similar *cis*-profiles (Hannenhalli and Levy, 2003). The proposed element-wise matrix multiplication can also be used to combine multiple types of data and reconstruct combination-specific gene networks by applying a in graph-theoretical approaches. The graph-theoretical analysis of human gene network obtained using the P*R association showed that it consists of a set of core nodes, mostly represented by genes involved in various signaling pathways, and numerous peripheral nodes represented by genes involved in oxidative phosphorylation. The analysis of human gene network obtained using P*R*K association showed that it is dominated by genes involved in extra-cellular interactions. Thus, different combinations of genome-wide data types reveal different types of gene networks.

Table4*. Pathways enriched in co-expressed genes.*

KEGG ID	z-score	p-value	Pathway name	KEGG category
hsa04080	9.73	0	Neuroactive ligand-receptor interaction	Environmental Information Processing
hsa04630	5.11	0	Jak-STAT signaling pathway	Environmental Information Processing
hsa04620	4.94	0.0003	Toll-like receptor signaling pathway	Cellular Processes
hsa00190	4.91	0.0003	Oxidative phosphorylation	Metabolism
hsa04020	4.69	0	Calcium signaling pathway	Environmental Information Processing
hsa00602	4.60	0.0007	Glycosphingolipid biosynthesis - neo-lactoseries	Metabolism
hsa04010	3.86	0.0006	MAPK signaling pathway	Environmental Information Processing
hsa04060	3.81	0.0005	Cytokine-cytokine receptor interaction	Environmental Information Processing
hsa04730	3.57	0.0022	Long-term depression	Cellular Processes
hsa04664	3.51	0.0022	Fc epsilon RI signaling pathway	Cellular Processes
hsa04320	3.50	0.0043	Dorso-ventral axis formation	Cellular Processes
hsa00534	3.40	0.0098	Heparan sulfate biosynthesis	Metabolism
hsa04140	3.04	0.0099	Regulation of autophagy	Genetic Information Processing
hsa04120	2.89	0.0094	Ubiquitin mediated proteolysis	Genetic Information Processing
hsa04742	2.87	0.0114	Taste transduction	Cellular Processes
hsa04330	2.85	0.0099	Notch signaling pathway	Environmental Information Processing
hsa04540	2.81	0.0082	Gap junction	Cellular Processes
hsa04910	2.76	0.0060	Insulin signaling pathway	Cellular Processes
hsa04370	2.72	0.0118	VEGF signaling pathway	Environmental Information Processing
hsa04740	2.69	0.0153	Olfactory transduction	Cellular Processes
hsa04650	2.66	0.0090	Natural killer cell mediated cytotoxicity	Cellular Processes
hsa00601	2.54	0.0299	Glycosphingolipid biosynthesis - lactoseries	Metabolism
hsa04610	2.50	0.0154	Complement and coagulation cascades	Cellular Processes
hsa00230	2.47	0.0150	Purine metabolism	Metabolism
hsa04930	2.46	0.0198	Type II diabetes mellitus	Human Diseases
hsa00510	2.45	0.0210	N-Glycan biosynthesis	Metabolism
hsa05120	2.37	0.0208	Epithelial cell signaling in Helicobacter pylori infection	Human Diseases
hsa04720	2.29	0.0231	Long-term potentiation	Cellular Processes
hsa04340	2.11	0.0312	Hedgehog signaling pathway	Environmental Information Processing
hsa04520	2.02	0.0350	Adherens junction	Cellular Processes
hsa04310	1.94	0.0388	Wnt signaling pathway	Environmental Information Processing
hsa04810	1.91	0.0362	Regulation of actin cytoskeleton	Cellular Processes
hsa04660	1.83	0.0450	T cell receptor signaling pathway	Cellular Processes
hsa04530	1.80	0.0455	Tight junction	Cellular Processes

***Table5**. Pathways enriched in both interacting and co-expressed genes.*

KEGG ID	z-score	p-value	Pathway name	KEGG category
hsa00193	78.58	0	ATP synthesis	Metabolism
hsa03050	76.25	0	Proteasome	Genetic Information Processing
hsa03020	44.87	0	RNA polymerase	Genetic Information Processing
hsa00190	36.83	0	Oxidative phosphorylation	Metabolism
hsa05110	30.64	0	Cholera	Human Diseases
hsa04350	29.41	0	TGF-beta signaling pathway	Environmental Information Processing
hsa04660	27.01	0	T cell receptor signaling pathway	Cellular Processes
hsa03010	25.52	0	Ribosome	Genetic Information Processing
hsa05120	25.18	0	Epithelial cell signaling in Helicobacter pylori infection	Human Diseases
hsa04610	23.94	0	Complement and coagulation cascades	Cellular Processes
hsa04620	23.80	0	Toll-like receptor signaling pathway	Cellular Processes
hsa00240	23.00	0	Pyrimidine metabolism	Metabolism
hsa04650	22.40	0	Natural killer cell mediated cytotoxicity	Cellular Processes
hsa03030	22.38	0	DNA polymerase	Genetic Information Processing
hsa04910	21.16	0	Insulin signaling pathway	Cellular Processes
hsa00230	20.90	0	Purine metabolism	Metabolism
hsa04320	20.85	0	Dorso-ventral axis formation	Cellular Processes
hsa04210	20.71	0	Apoptosis	Cellular Processes
hsa04010	20.54	0	MAPK signaling pathway	Environmental Information Processing
hsa00860	19.91	0	Porphyrin and chlorophyll metabolism	Metabolism
hsa04664	19.46	0	Fc epsilon RI signaling pathway	Cellular Processes

Another possible application of the proposed methodology is to benchmark the quality of various large-scale datasets. In this work, we used PPI from the OPHID database (Brown and Jurisica, 2005), where about 60% of all annotated interactions were inferred computationally, rather than obtained experimentally. Obviously, the quality of this inference needs to be validated. Since proteins that participate in the same functional pathway often form multi-protein complexes and can be expected to interact, the strength of K-P association can be used as an indicator of the non-randomness of PPI annotations. The fact that, according to our results, K-P association ranks highest among all pairwise combinations studied, suggests that the assignment of PPI in groups of functionally related proteins is highly non-random, thus confirming the quality of the OPHID annotation. Therefore, the present work can also be considered as an independent validation of OPHID, in addition to the validation provided by the authors of this database. A similar approach can be used to benchmark other types of data. For instance, given several methods for finding TFBS in promoter regions, the R-T association experiment can be used as a quantitative evaluation procedure to benchmark which of these methods gives the best correlation with expression data.

References

Alter O., Golub G.H. 2004. Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proc. Natl. Acad. Sci. USA* **101**:16577-16582.

Bader J.S., Chaudhuri A., Rothberg J.M., Chant J. 2004. Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* **22**:78-85.

Brown K.R., Jurisica I. 2005. Online predicted human interaction database. *Bioinformatics* **21**:2076-2082.

Carmona-Saez P., Chagoyen M., Rodriguez A., Trelles O., Carazo J.M., Pascual-Montano A. 2006. Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics* **7**:54.

Curwen V., Eyras E., Andrews T.D., Clarke L., Mongin E., Searle S.M., Clamp M. 2004. The Ensembl automatic gene annotation system. *Genome Res.* **14**:942-950.

Drawid A., Gerstein M. 2001. A Byesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome. *J. Mol. Biol.* **301**:1059-1075.

Garcia O., Saveanu C., Cline M., Fromont-Racine M., Jacquier A., Schwikowski B., Aittokallio T. 2007. GOlorize: A Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics* **23**:394-396.

Ge H., Liu Z., Church G.M., Vidal M. 2001. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **29**:482-486.

Gerstein M., Lan N., Jansen R. 2002. Proteomics. Integrating interactomes. *Science* **295**:284-287.

Goh C.S., Gianoulis T.A., Liu Y., Li J., Paccanaro A., Lussier Y.A., Gerstein M. 2006. Integration of curated databases to identify genotype-phenotype associations. *BMC Genomics* **7**:257.

Greenbaum D., Luscombe N.M., Jansen R., Qian J., Gerstein M., 2001. Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res.* **11:**1463-1468.

Grigoriev A. 2001. A relationship between gene expression and protein interactions on the proteome scale: Analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **29**:3513-3519.

Gunsalus K.C., Ge H., Schetter A.J., Goldberg D.S., Han J.D., Hao T., Berriz G.F., Bertin N., Huang J., Chuang L.S., Li N., Mani R., Hyman A.A., Sönnichsen B., Echeverri C.J., Roth F.P., Vidal M., Piano F. 2005. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**:861-865.

Hannenhalli S., Levy S. 2003. Transcriptional regulation of protein complexes and biological pathways. *Mamm. Genome* **14**:611-619.

Huang R., Wallqvist A., Covell D.G. 2006. Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen. *Genomics* **87**:315-328.

Hubbard T., Andrews D., Caccamo M., Cameron G., Chen Y., Clamp M., Clarke L., Coates G., Cox T., Cunningham F., Curwen V., Cutts T., Down T., Durbin R., Fernandez-Suarez X.M., Gilbert J., Hammond M., Herrero J., Hotz H., Howe K., Iyer V., Jekosch K., Kahari A., Kasprzyk A., Keefe D., Keenan S., Kokocinsci F., London D., Longden I., McVicker G., Melsopp C., Meidl P., Potter S., Proctor G., Rae M., Rios D., Schuster M., Searle S., Severin J., Slater G., Smedley D., Smith J., Spooner W., Stabenau A., Stalker J., Storey R., Trevanion S., Ureta-Vidal A., Vogel J., White S., Woodwark C., Birney E. 2005. Ensembl 2005. *Nucleic Acids Res.* **33**:D447-D453.

Hwang D., Rust A.G., Ramsey S., Smith J.J., Leslie D.M., Weston A.D., de Atauri P., Aitchison J.D., Hood L., Siegel A.F., Bolouri H. 2005. A data integration methodology for systems biology. *Proc. Natl. Acad. Sci. USA* **102**:17296-17301.

Jansen R., Greenbaum D., Gerstein M. 2002. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12**:37-46.

Kanehisa M., Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**:27-30.

Kel A.E., Gossling E., Reuter I., Cheremushkin E., Kel-Margoulis O.V., Wingender E. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31**:3576-3579.

Lee I., Date S.V., Adai A.T., Marcotte E.M. 2004. A probabilistic functional network of yeast genes. *Science* **306**:1555-1558.

Li S., Armstrong C.M., Bertin N., Ge H., Milstein S., Boxem M., Vidalain P.O., Han J.D., Chesneau A., Hao T., Goldberg D.S., Li N., Martinez M., Rual J.F., Lamesch P., Xu L., Tewari M., Wong S.L., Zhang L.V., Berriz G.F., Jacotot L., Vaglio P., Reboul J., Hirozane-Kishikawa T., Li Q., Gabel H.W., Elewa A., Baumgartner B., Rose D.J., Yu H., Bosak S., Sequerra R., Fraser A., Mango S.E., Saxton W.M., Strome S., Van Den Heuvel S., Piano F., Vandenhaute J., Sardet C., Gerstein M., Doucette-Stamm L., Gunsalus K.C., Harper J.W., Cusick M.E., Roth F.P., Hill D.E., Vidal M. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* **303**:540-543.

Myers C.L., Robson D., Wible A., Hibbs M.A., Chiriac C., Theesfeld C.L., Dolinski K., Troyanskaya O.G.

2005. Discovery of biological networks from diverse functional genomic data. *Genome Biology* **6**:R114.

Matys V., Fricke E., Geffers R., Gössling E., Haubrock M., Hehl R., Hornischer K., Karas D., Kel A.E., Kel-Margoulis O.V., Kloos D.U., Land S., Lewicki-Potapov B., Michael H., Münch R., Reuter I., Rotert S., Saxel H., Scheer M., Thiele S., Wingender E. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**:374-378.

Nagamine N., Kawada Y., Sakakibara Y. 2005. Identifying cooperative transcriptional regulations using protein-protein interactions. *Nucleic Acids Res.* **33**:4828-4837.

Qian J., Lin J., Luscombe N.M., Yu H., Gerstein M. 2003. Prediction of regulatory networks: Genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* **19**:1917-1926.

Robinson M., Sun Y., Boekhorst R.T., Kaye P., Adams R., Davey N., Rust A.G. 2006. Improving computational predictions of *cis*-regulatory binding sites. *Pac. Symp. Biocomput.* **11**:391-402.

Su A.I., Wiltshire T., Batalov S., Lapp H., Ching K.A., Block D., Zhang J., Soden R., Hayakawa M., Kreiman G., Cooke M.P., Walker J.R., Hogenesch J.B. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**:6062-6067.

Tanay A., Sharan R., Kupiec M., Shamir R. 2004. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. USA* **101**:2981-2986.

Vidal M. 2001. A biological atlas of functional maps. *Cell* **104**:333-339.

Wachi S., Yoneda K., Wu R. 2005. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**:4205-4208.

Walhout A.J., Reboul J., Shtanko O., Bertin N., Vaglio P., Ge H., Lee H., Doucette-Stamm L., Gunsalus K.C., Schetter A.J., Morton D.G., Kemphues K.J., Reinke V., Kim S.K., Piano F., Vidal M. 2002. Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* **12**:1952-1958.

Yu H., Luscombe N.M., Qian J., Gerstein M. 2003. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.* **19**:422-427.

Yang H.H., Hu Y., Buetow K.H., Lee M.P. 2004. A computational approach to measuring coherence of gene expression in pathways. *Genomics* **84**:211-217.

Zhang L.V., Wong S.L., King O.D., Roth F.P. 2004. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* **5**:38.