DRAFT – Please do not distribute

Implication Networks from Large Gene-expression Datasets

Debashis Sahoo[1], David L. Dill[2, 1], Andrew J. Gentles[4], Rob Tibshirani[3], Sylvia K. Plevritis[4]

[1]Department of Electrical Engineering, Stanford University, Stanford, CA, 94305, USA
[2]Department of Computer Science, Stanford University, CA, 94305, USA
[3]Department of Health Research and Policy and Department of Statistics, Stanford University, CA, 94305, USA
[4]Department of Radiology, Stanford University, CA, 94305, USA

Corresponding Author: David L. Dill (dill@cs.stanford.edu)
Phone: (650) 725-3642 Fax: (650) 725-6949

Total character count of the manuscript: 58,458

Keywords: Boolean Network/Microarray/Co-expression network/Pair-wise gene expression/Correlation/Co-regulation/Gene regulatory network

Subject categories: Bioinformatics, Computational methods

Implication Networks from Large Gene-expression Datasets

Abstract

We present a new algorithm for building Boolean networks from very large amounts of gene expression data. The resulting networks include not only symmetric relationships between genes, such as co-expression, but also asymmetric relations that represent if-then rules. The approach is conceptually simple and fast enough that it can build a complete gene network using 3 billion gene pairs with more than 9,500 expression values per gene-pair in less than 3 hours on an ordinary office computer. The algorithm was applied to publicly available data from thousands of microarrays for humans, mice, and fruit flies (for a total of 365 million Affymetrix probeset expression levels). The resulting network consists of hundreds of millions of relationships between genes, and contains biologically meaningful information about gender differences, tissue differences, development, differentiation and co-expression. We also examine relationships that are conserved between humans, mice, and fruit flies. The full Boolean relationships are available for exploration at http://gourd.stanford.edu/~sahoo/recomb07/.

Keywords: Boolean Network/Microarray/Co-expression network/Pair-wise gene expression/Correlation/Co-regulation/Gene regulatory network

Extended Synopsis

Introduction

A large and exponentially growing volume of gene expression data from microarrays is now available publicly. Since the quantity of data from around the world dwarfs the output of any individual laboratory, there are opportunities for data mining that can yield insights that would not be apparent from smaller, less diverse data sets. Consequently, there have been many efforts to extract large networks of relationships from large amounts of gene expression data. Almost all of this work constructs networks of pairwise relationships between genes, indicating that the genes are co-expressed (Allocco *et al.* 2004; Arkin and Ross 1995; Jordan *et al.* 2004; Lee *et al.* 2004; Tavazoie *et al.* 1999). Co-expression is a symmetric relationship (if A is related to B, then B is related to A), such as correlation.

This paper describes a new algorithm for building a Boolean network from large amounts of microarray data, and describes some of the properties of the resulting network. The algorithm classifies the expression level of each gene on each array as “low” or “high” relative to a threshold, and finds all Boolean relationships between pairs of genes, including not only symmetric relationships capturing co-expression, but also “if-then”

relationships (called implications) which are asymmetric. For example, a relationship could say "if gene A's expression level is high, then gene B's expression level is almost always low (more concisely, "A high implies B low" or "A high⇒B low"). It is important to understand that an implication does not necessarily imply causality. Instead, it is an empirically observed invariant on the expression levels of two genes. A network of implications is a *directed graph*, which connects nodes with arrows, instead of the more common *undirected graphs* used for many biological networks.

This approach is conceptually simple. The relationships are intuitive – they are immediately evident upon inspection of a scatterplot of the data points of expression levels for the two related genes, and are thus completely transparent to biologists, unlike some approaches, which find relationships that can be more difficult for users to interpret.

Symmetric relationships are implications in both directions. Genes A and B are strongly correlated, in general, when A high⇒B high and A low⇒B low. In this case, A and B are said to be *equivalent*. A second kind of symmetric relation occurs when A high⇒B low and A low⇒B high, the expression levels of A and B are usually strongly negatively correlated, and A and B are said to be *opposite*. Implications in one direction only are called *asymmetric*. Implications capture many more significant relations between pairs of genes than correlation. In other words, there may be a very significant Boolean implication between genes whose expression is very weakly correlated. There are six possible Boolean relationships: two symmetric (equivalent and opposite) and four asymmetric (low ⇒ low, low ⇒ high, high ⇒ low, high ⇒ high).

Boolean networks were constructed from 4,787 publicly available Affymetrix U133 Plus 2.0 human, 2,154 Affymetrix mouse 430 2.0, and 450 Affymetrix Drosophila genome 1 arrays from Gene Expression Omnibus (Edgar *et al.* 2002). All the datasets were normalized using the RMA algorithm (Irizarry *et al.* 2003). There are 208 million, 336 million and 17 million Boolean relationships in human, mouse and fruit fly respectively. Additionally, 4 million Boolean relationships are conserved in human and mouse and 41,260 Boolean relationships are conserved in human, mouse and fruit fly. The algorithm is fast enough to scale to large volumes of data. The algorithm that builds a complete gene network using 54,677x54,677 gene pairs with more than 9,500 expression values per gene-pair in less than 3 hours. The Boolean relationships are available for exploration at http://gourd.stanford.edu/~sahoo/recomb07/.

The relationships in the resulting network are often biological meaningful. Differences associated with gender and tissue-type is readily apparent. Relationships between genes that are active only during specific developmental or differentiation stages are also evident. Large groups of equivalent genes associated with the cell cycle appear in the network for each species. Highly conserved relationships are enriched with the cell cycle and central nervous system specific genes.

Results

Boolean relationships are present in gene expression microarray data.

The two symmetric Boolean relationships correspond to two sparse quadrants in a scatterplot, as described in the materials and methods section. First, the low-high and high-low quadrant can be sparse as shown in Figure 1(a), which shows that CCNB2 and BUB1B are equivalent. Highly positively correlated genes are almost always equivalent. Alternatively, the low-low and high-high quadrants can be sparse, as shown in Figure 1(d), which shows that EED and XTP7 are opposite.. Negatively correlated genes are often opposite. Notice that it is not possible to have both the low-low and high-low quadrants be sparse because that would require the second gene to be always low; similarly, it is not possible for the low-high and low-low quadrants both be sparse.

There are four possible asymmetric Boolean relationships,, which correspond to one sparse quadrant. Figure 1(b) shows where the quadrant for gene PTPRC low and gene CD19 high is sparse so PTPRC low⇒CD19 low. Figure 1(c) shows that XIST high ⇒RPS4Y1 low (this relationship was previously pointed out in paper describing the CELSIUS database of microarray data (Day *et al.* 2007), while annotating microarrays with male and female). Figure 1(e) shows FAM60A low⇒NUAK1 high. In this case, when FAM60A expression level is low, NUAK1 expression level is high, but when FAM60A expression level is high, NUAK1 expression level is evenly distributed between high and low. Finally, Figure 1(f) shows that COL3A1 high⇒SPARC high. This relationship is complex, since it can be viewed as a combination of multiple kinds of relationships including linear and constant. However, Boolean analysis discovers the simple logical implication: COL3A1 high⇒SPARC high.

Notice that for each of the above Boolean relationships there is always a *contrapositive* Boolean relationship that holds. For example, PTPRC low⇒CD19 low so CD19 high⇒ PTPRC high. Similarly, XIST high⇒RPS4Y1 low, so RPS4Y1 high⇒XIST low, FAM60A low⇒NUAK1 high so NUAK1 low⇒FAM60A high and COL3A1 high⇒ SPARC high so SPARC low ⇒ COL3A1 low.

A large number of Boolean relationships exist in gene expression data.

A very large number of Boolean relationships were found in microarray data for individual species. There are 208 million implications in the human dataset, even with a stringent requirement for significance (a permutation test yields a false discovery rate (FDR) of 10^{-4}). The mouse dataset has 336 million implications (FDR = $6x10^{-5}$), and the fruit fly dataset has 17 million implications (FDR = $6x10^{-6}$). Of the 208 million implications in the human dataset, 128 million are high⇒low, 38 million are low⇒low, 38 million are high⇒high, 2 million are low⇒high, 1.6 million relations are equivalences and 0.4 million are opposite. Table 1 summarizes the number of Boolean relationships found in each dataset. In all cases, the most common relationships are the high⇒low type, and the opposite relations are the most uncommon. As can be seen from Table 1, in the human dataset 1% of the total Boolean relationships are symmetric, while the

remaining 99% are asymmetric. Similarly, in the mouse dataset 1.4% of the total Boolean relationships are symmetric, and 98.6% are asymmetric. However, in fruit fly 12% of the Boolean relationships are symmetric. The number of low⇒low relationships is the same as the number of high⇒high relationships because of contrapositives.

Asymmetric Boolean relationships are far more numerous than symmetric relationships.

Discovery of asymmetric Boolean relationships is one of the novelties of Boolean analysis, as they have not been explored in the literature thoroughly. Our analysis discovers a large number of asymmetric Boolean relationships (low⇒high, high⇒low, low⇒low and high⇒high) compared to symmetric Boolean relationships (equivalent and opposite).

Networks based on correlation of gene expression would fail to include these asymmetric relations. 98.8% of the asymmetric Boolean relationships on the human CD genes have correlation coefficients ranging from –0.65 to 0.65. Further, as expected most of the low⇒high and high⇒low relationships have negative correlation coefficients. The low⇒high relationships have correlation coefficients from –0.55 to 0 and the high⇒low relationships have correlation coefficients from –0.65 to 0.25 as shown in Figure 2(f) and Figure 2(c) respectively.

Low⇒low and high⇒high have mostly positive correlation coefficients, from –0.15 to 0.95, as shown in Figure 2(b) and Figure 2(g). (They have exactly same distribution of correlation coefficients because of contrapositives.) Some of these relationships have very high correlation coefficients; for example, relationships with correlation coefficient 0.933 and 0.7963 are shown in Figure 2(h) and Figure 2(i).

Boolean equivalences compares well to correlation-based approaches.

In order to compare the properties of Boolean networks to more common correlation-based networks, both types of networks were constructed based on human CD antigen genes. These genes were chosen as a relatively small and coherent subset of biologically interesting genes. A complete correlation-based network on human CD genes was computed as described in Materials and Methods.

Figure 2 shows histograms of the Boolean relationships with respect to the Pearson's correlation coefficients. As can be seen from the figure, the number of equivalences from the Boolean network that are also in the correlation network increases linearly with the correlation coefficient. Gene pairs that have no Boolean relationships also have low correlation coefficients around zero. There are a substantial number of Boolean relationships, for which the correlation coefficient is small. These pairings cannot be identified by a pure correlation-based approach. Four examples of scatter plots are shown in Figure 2 (bottom row) to demonstrate the differences between Boolean relationships and correlation-based relationships.

DRAFT – Please do not distribute

Boolean networks are not scale free.

It has often been observed that other biological networks are scale-free (Barabasi and Albert 1999; Barabasi and Oltvai 2004; Bhan *et al.* 2002; Featherstone and Broadie 2002; Jeong *et al.* 2000; Jeong *et al.* 2001), to study the global properties of Boolean network, we plotted frequency of the probesets against their degree (number of Boolean relationships) as shown in Figure 3. Each log-log plot shows on the horizontal axis the degree, while the vertical axis shows the number of probesets that have the number of relationships to other probesets. The top row in Figure 3 corresponds to the human Boolean network. From left to right are shown the total Boolean relationships, only symmetric Boolean relationships, and only asymmetric Boolean relationships. These plots are comparable to the Boolean networks for mouse and fruit fly (as shown in Supplementary Figure 1). The middle row in Figure 3 corresponds to the conserved Boolean network between human and mouse, constructed of relationships that are present in both human and mouse. Finally, the bottom row in Figure 3 shows the conserved Boolean network between human, mouse and fruit fly. As can be seen from the figures, the plots for total Boolean relationships (1st column in Figure 3) are non-linear. However, the plots for symmetric and asymmetric Boolean relationships (2nd and 3rd columns in Figure 3) are close to linear. Interestingly, although one might anticipate that a network, which has not been transitively reduced (see materials and methods), would have more nodes with high degree (relative to a power law behavior), we found the opposite. The Boolean network described here scales below a power law at high degree.

Boolean relationships are highly conserved across species.

A network can be constructed consisting of the relations that hold between orthologous genes in multiple species. The network of relationships that are conserved in humans and mice network has a total of 3.2 million Boolean relationships consisting of 8,000 low⇒high, 2 million high⇒low, 0.5 million low⇒low, 0.5 million high⇒high, 10,814 equivalent and 94 opposite implications. Applying the same analysis to randomized human and mouse datasets yielded *no* conserved Boolean relationships, for an estimated false discovery rate of less than 3.1e-7. An analogous network of implications conserved across human, mouse and fruit fly has 41,260 Boolean relationships: 24,544 high⇒low, 8,060 low⇒low, 8,060 high⇒high and 596 equivalent. The false discovery rate for the conserved human, mouse and fruit fly Boolean network is less than 2.4e-5.
Figure 4 shows three examples of highly conserved Boolean relationships from human, mouse and fruit fly. The first row in Figure 4 is an example of equivalent relationships that are conserved in all three species. The middle and bottom rows show highly conserved high⇒low and high⇒high relationships.

The connected components of the network of equivalent relationships that were conserved in human, mouse, and fruit fly were examined (a connected component of an undirected graph is a set of genes where there is a path between every pair of genes). The algorithm found 13 different connected components. However, there are two distinct large components. The largest component has 178 genes including BUB1B, EZH2, CCNA2, CCNB2 and FEN1. The genes belong to this component were analyzed using

DAVID functional annotation tools (Dennis *et al.* 2003; Hosack *et al.* 2003). The functional annotation analysis indicates DNA replication (2.03e-14, 19 genes) and cell cycle process (1.06e-13, 30 genes) as significant interesting gene ontology annotations for the largest component. The second component has 32 genes with transport (2.55e-08, 16 genes) and synaptic transmission (1.04e-08, 8 genes) as significant gene ontology annotations. Further, the functional annotation analysis discovers proteasome and cell cycle as significant KEGG pathways for the first component. The second component was enriched for calcium signaling pathway in KEGG database. The list of genes for the components and the DAVID functional annotation results are included in the supplementary information.

Boolean network computation is fast.

The total computation time to construct the network of implications for the human dataset was 2.5 hours on a 2Ghz computer with 8GB of memory. The human dataset consisted of a total of 54,677 distinct probesets from 4,787 microarrays. The computation time for the mouse dataset was 1.6 hours. This data set has 45,101 probesets and 2,154 microarrays. Finally, the computation time for fruit fly dataset, consisting of 14,010 probesets and 450 microarrays, was 2 minutes.

Discussion

Boolean analysis is simple, fast and efficient.

Boolean analysis provides a simple intuitive characterization of relationships between pairs of genes. A threshold is determined to classify low and high values, after fitting a step function to the sorted gene expression levels using the StepMiner algorithm (Sahoo *et al.* 2007), which sets the threshold near the mean of uniformly distributed sets, and otherwise places it at the largest gap between clusters of relatively low and high values. The Boolean analysis algorithm finds a large number of Boolean implications even if the gene expression values are not Boolean. In many cases the relationships are more complicated than Boolean implication. However, it may be useful to view them as Boolean implications because these are easy to manipulate. Also, Boolean relationships are fast to compute because bit vector operations are fast compared to floating point operations. The computation time for the human dataset was 2.5 hours, whereas correlation coefficient computation might take multiple days in a single computer.

Boolean equivalence vs. correlation-based relationship

As shown in Figure 2, gene pairs with high correlation coefficient are more likely to be equivalent in the Boolean analysis. However, there are exceptions, as shown in the plot between COL3A1 and COL1A1. Here, the correlation coefficient (0.933) is extremely high and the relationship is expected to be equivalent. However, there are many microarrays where the expression levels for COL3A1 is high and COL1A1 is low.

Therefore, Boolean analysis concludes this relationship as COL3A1 low ⇒ COL1A1 low. However, there is a very strong linear component in the scatter plot, which a correlation-based relationship can distinguish. Both the Boolean and the correlation-based characterization may be important biologically in different contexts. Most methods to compute correlation-based networks use a threshold of more than 0.7 on the correlation coefficient (Jordan *et al.* 2004; Lee *et al.* 2004; Tsaparas *et al.* 2006). Figure 2 shows that a very large proportions of asymmetric Boolean relationships have correlation coefficient less than 0.65. Note that 20% of the symmetric Boolean relationships have correlation coefficient less than 0.65. Correlation-based networks will miss these relationships, which might be biologically relevant. Figure 2 shows an example scatter plot between LAIR1 and WAS with correlation coefficient 0.5158 that our approach identifies as an equivalence. Similarly, TLR2 and ITGAM have a correlation coefficient of 0.7 and are considered equivalent. However, the plot between VPREB1 and IGLL1 shows a higher correlation coefficient than 0.7, and we infer an asymmetric Boolean relationship (VPREB1 high ⇒ IGLL1 high).

Asymmetric Boolean relationships

As can be seen, asymmetric Boolean relationships are prevalent in the Boolean analysis. Moreover, a huge percentage of these relationships are high⇒low. One can imagine that if there were *n* mutually exclusively expressed genes, there would be *n**(*n*-1) high⇒low relationships. Furthermore, tissue specific genes are often mutually exclusively expressed and could be a major contributor of the high⇒low relationships. Additionally, asymmetric relationships have low correlation coefficient, as expected because Pearson's correlation is symmetric, suggesting that correlation-based approaches identify them poorly. Interestingly, although one might imagine that different probesets for the same gene should have positive symmetric relationships, we find that they have asymmetric Boolean relationships consistent with previous findings of low average correlation among them (Liao and Zhang 2006). Therefore, Boolean analysis might be helpful in pointing out important differences among different probesets for the same gene. We believe that asymmetric Boolean relationships are rich in important biological relationships and might be helpful in generating new biological hypotheses in the future. We have shown that large numbers of asymmetric relationships are highly conserved and they follow some of the currently known biological phenomena.

Highly conserved Boolean relationships

A conserved Boolean relationship is one that holds between orthologous genes across diverse species. There are many symmetric and asymmetric relationships that are conserved in human, mouse and fruit fly. Figure 4 shows three examples. The top row in Figure 4 shows that CCNB2 (CycB in fruitfly) and BUB1B are equivalent in all three species. (In this case, a network of correlated genes would also be able to find these conserved relationships because they are very well correlated in each species.). Finding this relationship becomes feasible because of the efficiency of the Boolean analysis.
It is very well known that both CCNB2 and BUB1B are related to cell cycle (Bolognese *et al.* 1999; Davenport *et al.* 1999). It might not be surprising to see that they are very

highly correlated in all three species. However, it is surprising to note that only a small number of currently known cell cycle genes have this property. The bottom row in Figure 4 shows an asymmetric relationship between two very well known cell cycle regulators, E2F2 and PCNA (Ivey-Hoyle *et al.* 1993; Mathews *et al.* 1984; Miyachi *et al.* 1978). The middle row in Figure 4 shows an asymmetric relationship that is conserved in all three species. GABRB1 is a receptor to an inhibitory neurotransmitter in vertebrate brain (Kirkness *et al.* 1991). It is surprising to see that the relationship between GABRB1 and BUB1B is conserved in vertebrate and arthropods (fruit fly). This relationship might suggest that cells expressing this particular neurotransmitter are less likely to be proliferating. To our knowledge, this is the first algorithm that finds asymmetric Boolean relationships that are conserved across species as diverse as vertebrates and arthropods.

Boolean relationships show gender differences, tissue differences, development, differentiation and co-expression.

Boolean relationships represent a wide variety of currently known biological phenomena. The generated networks contain relationships that show gender differences, development, differentiation, tissue difference and co-expression. The scatter plot between XIST and RPS4Y1 in Figure 5(a) is an example of an asymmetric Boolean relationship that shows gender difference. RPS4Y1 is expressed only in certain male tissues because it is present solely on the Y chromosome (Weller *et al.* 1995) and XIST is normally expressed only in female tissues (Brockdorff *et al.* 1991; Brown *et al.* 1991), so RPS4Y1 and XIST are rarely expressed together on the same array. Hence, the relationships RPS4Y1 high $\Rightarrow$ XIST low and XIST high $\Rightarrow$ RPS4Y1 low hold. In the network, RPS4Y1 is equivalent to four other genes, all of which are Y-linked. RPS4Y1 low$\Rightarrow$ ACPP low (Figure 5(b)), KLK2 low, and KLK3 (PSA) low, all of which are prostate-specific (Sharief *et al.* 1994). Some of the relationships capture the hierarchy of tissue types. For example, GABRB1 is specific to the central nervous system (Roth *et al.* 2006), and ACPP high $\Rightarrow$ GABRB1 low (Figure 5(c)), because the prostate is distinct from the CNS. On the other hand, GABRA6 is primarily expressed in the cerebellum, and we see that GABRB1 low $\Rightarrow$ GABRA6 low, because the cerebellum is part of the CNS (more literally, if a tissue sample is not part of the CNS, it is also not part of the cerebellum).

To show an example of a Boolean relationship between two developmentally regulated genes, we identify HOXD3 and HOXA13 as shown in Figure 5(d). HOXD3 and HOXA13 have their evolutionary origin from fruit fly antennapedia (Antp) and ultrabithorax (UBX) respectively (Carroll 1995). It was recently discovered that HOXD3 and HOXA13 are expressed in human proximal and distal sites respectively (Rinn *et al.* 2007), a pattern of expression, which is evolutionarily conserved from fruit flies. The human Boolean network indeed shows that high expression of HOXD3 and HOXA13 are mutually exclusive (HOXD3 high $\Rightarrow$ HOXA13 low), which is consistent with the above paper. (Contrary to the findings of that paper, this relationship is not highly conserved in our analysis because the mouse and fruit fly orthologous probes for the desired genes did not have a good dynamic range in the dataset, for unknown reasons.)

Relationships between genes expressed during the process of differentiation also appear in the network. For example, a Boolean relationship between two key marker genes from B cell differentiation, KIT and CD19 as shown in Figure 5(e). KIT is a hematopoietic stem cell marker (Ikuta *et al.* 1991) and CD19 is a well-known B cell differentiation marker (Stamenkovic and Seed 1988). Our algorithm discovers that KIT and CD19 are rarely expressed together and thus the Boolean relationship: CD19 high $\Rightarrow$ KIT low and KIT high $\Rightarrow$ CD19 low.

From inspecting the human network, it is clear that hundreds of genes are co-expressed that are related to the cell cycle. Two such genes, CDC2 and CCNB2, are shown in Figure 5(f).

The Boolean network is not scale free

As shown in Figure 3, the log-log plots between degree of connectedness, and frequency is highly non-linear. For a scale free network we expect the plot to show a linear power law distribution (Barabasi and Albert 1999; Barabasi and Oltvai 2004; Bhan *et al.* 2002; Featherstone and Broadie 2002; Jeong *et al.* 2000; Jeong *et al.* 2001). It has been shown (Barabasi and Albert 1999) that the combination of continual additions of nodes to a network, together with the property that new connections are more likely to be made to a highly connected node, naturally leads to power law behavior. However, our Boolean analysis found sub-power law scaling of number of nodes with respect to their degree.

Comparison with other approaches for building gene networks

Traditional analysis of a large microarray dataset often begins with pairwise analysis of genes. A large number of algorithms have been proposed to infer biologically relevant gene pairs, presented in the form of a gene regulatory network or a co-expression network (Allocco *et al.* 2004; Arkin and Ross 1995; Jordan *et al.* 2004; Lee *et al.* 2004; Tavazoie *et al.* 1999). Most clustering algorithms (Cho *et al.* 1998; Eisen *et al.* 1998; Spellman *et al.* 1998) also rely on pairwise gene expression analysis. Sophisticated algorithms including, Bayesian analysis (Friedman *et al.* 2000; Friedman 2004; Lee *et al.* 2006; Li and Chan 2004; Pe'er *et al.* 2001; Segal *et al.* 2004; Segal *et al.* 2005; Segal *et al.* 2001), Graphical Gaussian Models (Kishino and Waddell 2000; Schafer and Strimmer 2005) and mutual information (Basso *et al.* 2005; Butte and Kohane 2000; Margolin *et al.* 2006; Wang *et al.* 2005) have been employed to infer the underlying network structure. Most of the above approaches discover symmetric relationships and require pairwise gene expression analysis. The main drawback of the pairwise gene expression analysis is the computation time required to investigate a large number of gene pairs. A massively parallel grid-computing environment has been used to reduce the computation time (Swain *et al.* 2005), but this approach demands costly machines. Boolean analysis finds a large number of asymmetric relationships and it is relatively fast compared to most of the above approaches.

Previously developed Boolean networks have only been applied to smaller sized datasets (Gupta *et al.* 2007; Ideker *et al.* 2000; Kauffman 1971; Liang *et al.* 1998; Pal *et al.* 2005;

Shmulevich and Zhang 2002; Shmulevich and Kauffman 2004). Further, Boolean implication network similar to our network have been used for probabilistic reasoning (Liu and Desmarais 1997).

Large sized microarray datasets have been collected, analyzed (Day *et al.* 2007; Rhodes *et al.* 2007) and applied to the study of human cancer (Hanauer *et al.* 2007). Boolean analysis can potentially be applied to the above datasets to explore meaningful Boolean relationships. Recently, in the CELSIUS database (Day *et al.* 2007), a gene coexpression network was built using a subset of 3600 probesets only. However, it is feasible to apply our Boolean analysis to the full dataset.

Comparison with previous approaches for building conserved gene-interaction network

Conservation across multiple species has been used to infer more likely regulatory relationships (Chalmel *et al.* 2007; Sinha *et al.* 2004; Strand *et al.* 2007; Stuart *et al.* 2003; Tamada *et al.* 2005; Tirosh *et al.* 2006; Tsaparas *et al.* 2006; van Noort *et al.* 2003). Many of these algorithms suffer from the same computational bottleneck as building co-expression networks or clustering. Moreover, it is hard to detect conserved pairs with low correlation coefficient, although they too may be biologically meaningful, as we will demonstrate. It is easy to perform conservation analysis on Boolean network, which involves checking if the orthologous gene pairs have the same Boolean relationships, while other approaches require non-trivial probabilistic measure of conservation. Numerous studies use co-expression, while building conserved gene-interaction networks. An early study of this type (van Noort *et al.* 2003) improved the accuracy of predicting functional gene interactions by using conserved co-expression between *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. They used a correlation coefficient threshold of 0.6. Subsequently, another study (Stuart *et al.* 2003) identified 22,163 gene pairs from 3,182 DNA microarrays from humans, flies, worms and yeast. This study used a rank order statistic to compute a probabilistic measure of the conserved coexpression in multiple species. Further, Bayesian analysis was combined with conservation to build gene networks for yeast and human using cell cycle data (Tamada *et al.* 2005). Later studies focus on human and mouse to discover conserved gene expression in brain (Strand *et al.* 2007) and gametogenesis (Chalmel *et al.* 2007). None of the above algorithms predicts conserved asymmetric relationships. In addition to symmetric relationships (equivalent and opposite), Boolean analysis is also capable of discovering a large number of conserved asymmetric Boolean relationships. Moreover, Boolean analysis provides more transparent pair-wise relationships (as the Boolean relationships are directly visible in the scatter plot) compared to other approaches described above.

Materials and methods

Data collection and preprocessing

CEL files for 4,787 Affymetrix U133Plus 2.0 human microarrays, 2,154 Affymetrix 430 2.0 mouse arrays, and 450 Affymetrix Genome 1.0 Drosophila were downloaded from NCBI's Gene Expression Omnibus (Edgar *et al.* 2002). These array types were chosen because they are widely used, and because results from different arrays can be compared more easily than results from two-channel arrays. The datasets were normalized using the standard RMA algorithm (Irizarry *et al.* 2003); however, the available version of RMA uses excessive amounts of primary memory when normalizing thousands of arrays, so the program was re-written to increase memory efficiency. Boolean expression levels were assigned for each gene in each array, using the log (base 2) of the expression values (Figure 6 illustrates this process). First, a threshold was assigned to each gene using the StepMiner algorithm (Sahoo *et al.* 2007), which was originally designed to fit step functions to time-course data. For this application, the expression values for each gene were ordered from low-to-high, and StepMiner was used to fit a rising step function to the data that minimizes the differences between the fitted and measured values. This approach places the step at the largest jump from low values to high values (but only if there are sufficiently many expression values on each side of the jump to provide evidence that the jump is not due to noise), and sets the threshold at the point where the step crosses that original data (as shown in Figure 6). In the case where the gene expression levels are evenly distributed from low to high, the threshold tends to be near the mean expression level. If the assigned threshold for a gene is t, expression levels above $t + 0.5$ are classified as "high," expression levels below $t–0.5$ are classified as "low," and values between $t –0.5$ and $t+0.5$ are classified as "intermediate." Whenever more than 2/3 of the expression values of a gene were at an intermediate level of expression, the gene was excluded from further analysis, due to insufficient dynamic range in the expression values.

Discovery of Boolean relationships

All pairs of features with sufficient dynamic range were analyzed to discover potential Boolean relationships. There are six possible Boolean relationships between genes A and B that are constructed from four possible Boolean implications: A low $\Rightarrow$ B low, A low $\Rightarrow$ B high, A high $\Rightarrow$ B low, and A high $\Rightarrow$ B high. Each of the above implication is detected by checking whether one of the four quadrants in the scatter plot of Figure 6 is significantly sparsely populated with points compared with the other quadrants (intermediate values for A and B are ignored in this analysis). There are at most two possible sparse quadrants because the thresholds always separate a reasonable number of low and high expression levels for each gene. Each sparse quadrant corresponds to an implication. If A high $\Rightarrow$ B high and A low $\Rightarrow$ B low, A and B are considered to have equivalent levels of Boolean expression. When A high $\Rightarrow$ B low and A low $\Rightarrow$ B high, A and B are considered to have an opposite Boolean relationship. In both of these cases, two diagonally opposite quadrants are significantly sparse. In other cases, where there is only one sparse quadrant, the Boolean relationships between A and B have the same

name as Boolean implications: A low ⇒ B low, A low ⇒ B high, A high ⇒ B low, and A high ⇒ B high. There are two tests that must succeed for the relationship between A and B to be considered an implication. The following tests are performed to check whether the low-low quadrant is sparse that gives A low ⇒ B high. First, the number of expression values in the sparse quadrant must be significantly less than the number that would be expected under an independence model, given the relative distribution of low and high values for A and B. Specifically, if a_{00}, a_{01}, a_{10}, a_{11} are the number of expression values where A and B are low and low, low and high, high and low, and high and high, respectively, a threshold on the following statistic is performed to test whether the low-low quadrant is sparse.

$$total = a_{00} + a_{01} + a_{10} + a_{11}$$
$$expected = (a_{00} + a_{01}) * (a_{00} + a_{10}) / total$$
$$observed = a_{00}$$
$$statistic = \frac{(expected - observed)}{\sqrt{expected}}$$

Second, the observed values in the sparse quadrant are considered erroneous points and a sparse quadrant must have a small number of erroneous points. A maximum likelihood estimate of the *error rate* is computed as follows.

$$error\ rate = \frac{1}{2}\left(\frac{a_{00}}{(a_{00} + a_{01})} + \frac{a_{00}}{(a_{00} + a_{10})}\right)$$

A second threshold on this error rate is performed to ensure that the quadrant is really sparse. If the above tests succeed, the low-low quadrant is considered sparse and therefore, A low ⇒ B high is inferred. Similarly, the above tests are repeated for all other quadrants. A threshold of 3 for the first *statistic* and a threshold of 0.1 for the *error rate* are used here to discover the Boolean relationships. A Boolean network (directed graph) is built from the Boolean relationships, where nodes are A high or A low for each gene A and edges are Boolean implications. For example, there is a directed edge from A low to B high if there is a Boolean implication A low ⇒ B high. Boolean implications are transitive *e.g.* whenever A low ⇒ B high and B high ⇒ C low, A low ⇒ C low. Therefore, the Boolean network is almost transitively closed. A straightforward transitive reduction algorithm, however, can be applied to reduce the network size.

Computation of False Discovery Rate

To compute the false discovery rate (Storey and Tibshirani 2003), we permute randomly the expression values for each gene independently. Then, we run the Boolean analysis described above to build a complete Boolean network. The above analysis is repeated twenty times to compute the average number of Boolean relationships in the randomized data. The ratio of the average number of Boolean relationships in the randomized data to the original data is considered the false discovery rate of the Boolean analysis.

DRAFT – Please do not distribute

Correlation network for human CD genes

Human CD (cluster of differentiation) genes were selected for comparison against a correlation-based network. The set of genes includes 966 Affymetrix U133 Plus 2.0 human probesets. Pearson's correlation coefficients for all 466,095 pairs of genes were computed. Boolean analysis is also performed on this data to compare Boolean network with the correlation-based network.

Discovery of conserved Boolean relationships

Mouse and fruit fly orthologs for human genes were selected from the EUGene database (Gilbert 2002). For each Boolean relationship in the human dataset, a conserved relationship is detected if any of the mouse orthologs of the first human gene has a significant Boolean relationship with another mouse ortholog of the second human gene. To find conserved Boolean relationships in all three species, we check if any of the fruit fly orthologs of the first mouse gene has a significant Boolean relationship with another fruit fly orthologs of the second mouse gene for each conserved relationships in human and mouse.

Connected component analysis

Human genes for the highly conserved relationships in all three species were selected for the connected component analysis. An undirected graph was built with the gene names as nodes and the edges are from Boolean equivalent relationships. Connected component analysis was performed using a standard union-find algorithm on the undirected graph to find clusters of genes that are connected together.

Acknowledgements

The authors would like to thank NIH for supporting this work through Grant 5U56CA112973-02.

References

Allocco DJ, Kohane IS, Butte AJ. 2004. Quantifying the relationship between co-expression, co-regulation and gene function. BMC bioinformatics 5:18.

Arkin A and Ross J. 1995. Statistical construction of chemical reaction mechanisms from measured time-series. J. Phys. Chem. :970-979.

Barabasi AL and Albert R. 1999. Emergence of scaling in random networks. Science 286:509-512.

Barabasi AL and Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. Nature reviews.Genetics 5:101-113.

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. 2005. Reverse engineering of regulatory networks in human B cells. Nature genetics 37:382-390.

Bhan A, Galas DJ, Dewey TG. 2002. A duplication growth model of gene expression networks. Bioinformatics (Oxford, England) 18:1486-1493.

Bolognese F, Wasner M, Dohna CL, Gurtner A, Ronchi A, Muller H, Manni I, Mossner J, Piaggio G, Mantovani R, Engeland K. 1999. The cyclin B2 promoter depends on NF-Y, a trimer whose CCAAT-binding activity is cell-cycle regulated. Oncogene 18:1845-1853.

Brockdorff N, Ashworth A, Kay GF, Cooper P, Smith S, McCabe VM, Norris DP, Penny GD, Patel D, Rastan S. 1991. Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. Nature 351:329-331.

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, Willard HF. 1991. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature 349:38-44.

Butte AJ and Kohane IS. 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pacific Symposium on Biocomputing.Pacific Symposium on Biocomputing :418-429.

Carroll SB. 1995. Homeotic genes and the evolution of arthropods and chordates. Nature 376:479-485.

Chalmel F, Rolland AD, Niederhauser-Wiederkehr C, Chung SS, Demougin P, Gattiker A, Moore J, Patard JJ, Wolgemuth DJ, Jegou B, Primig M. 2007. The conserved transcriptome in human and rodent male gametogenesis. Proceedings of the National Academy of Sciences of the United States of America 104:8346-8351.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular cell 2:65-73.

Davenport JW, Fernandes ER, Harris LD, Neale GA, Goorha R. 1999. The mouse mitotic checkpoint gene bub1b, a novel bub1 family member, is expressed in a cell cycle-dependent manner. Genomics 55:113-117.

Day A, Carlson MR, Dong J, O'connor BD, Nelson SF. 2007. Celsius: a community resource for Affymetrix microarray data. Genome biology 8:R112.

Dennis G,Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology 4:P3.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research 30:207-210.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 95:14863-14868.

Featherstone DE and Broadie K. 2002. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. BioEssays : news and reviews in molecular, cellular and developmental biology 24:267-274.

Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. Science (New York, N.Y.) 303:799-805.

Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian networks to analyze expression data. Journal of computational biology : a journal of computational molecular cell biology 7:601-620.

Gilbert DG. 2002. euGenes: a eukaryote genome information system. Nucleic acids research 30:145-148.

Gupta S, Bisht SS, Kukreti R, Jain S, Brahmachari SK. 2007. Boolean network analysis of a neurotransmitter signaling pathway. Journal of theoretical biology 244:463-469.

Hanauer DA, Rhodes DR, Sinha-Kumar C, Chinnaiyan AM. 2007. Bioinformatics approaches in the study of cancer. Current Molecular Medicine 7:133-141.

Hosack DA, Dennis G,Jr, Sherman BT, Lane HC, Lempicki RA. 2003. Identifying biological themes within lists of genes with EASE. Genome biology 4:R70.

Ideker TE, Thorsson V, Karp RM. 2000. Discovery of regulatory interactions through perturbation: inference and experimental design. Pacific Symposium on Biocomputing.Pacific Symposium on Biocomputing :305-316.

Ikuta K, Ingolia DE, Friedman J, Heimfeld S, Weissman IL. 1991. Mouse hematopoietic stem cells and the interaction of c-kit receptor and steel factor. International journal of cell cloning 9:451-460.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic acids research 31:e15.

Ivey-Hoyle M, Conroy R, Huber HE, Goodhart PJ, Oliff A, Heimbrook DC. 1993. Cloning and characterization of E2F-2, a novel protein with the biochemical properties of transcription factor E2F. Molecular and cellular biology 13:7802-7812.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41-42.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. 2000. The large-scale organization of metabolic networks. Nature 407:651-654.

Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV. 2004. Conservation and coevolution in the scale-free human gene coexpression network. Molecular biology and evolution 21:2058-2070.

Kauffman S. 1971. Gene regulation networks: a theory for their global structure and behaviors. Current topics in developmental biology 6:145-182.

Kirkness EF, Kusiak JW, Fleming JT, Menninger J, Gocayne JD, Ward DC, Venter JC. 1991. Isolation, characterization, and localization of human genomic DNA encoding the beta 1 subunit of the GABAA receptor (GABRB1). Genomics 10:985-995.

Kishino H and Waddell PJ. 2000. Correspondence analysis of genes and tissue types and finding genetic links from microarray data. Genome informatics.Workshop on Genome Informatics 11:83-95.

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. 2004. Coexpression analysis of human genes across many microarray data sets. Genome research 14:1085-1094.

Lee SI, Pe'er D, Dudley AM, Church GM, Koller D. 2006. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proceedings of the National Academy of Sciences of the United States of America 103:14062-14067.

Li Z and Chan C. 2004. Inferring pathways and networks with a Bayesian framework. The FASEB journal : official publication of the Federation of American Societies for Experimental Biology 18:746-748.

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

Liao BY and Zhang J. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Molecular biology and evolution 23:530-540.

Liu J and Desmarais M. 1997. A Method of Learning Implication Networks from Empirical Data: Algorithm and Monte-Carlo Simulation-Based Validation. Knowledge and Data Engineering 9:990-1004.

Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A. 2006. Reverse engineering cellular networks. Nature protocols 1:662-671.

Mathews MB, Bernstein RM, Franza BR,Jr, Garrels JI. 1984. Identity of the proliferating cell nuclear antigen and cyclin. Nature 309:374-376.

Miyachi K, Fritzler MJ, Tan EM. 1978. Autoantibody to a nuclear antigen in proliferating cells. Journal of immunology (Baltimore, Md.: 1950) 121:2228-2234.

Pal R, Datta A, Fornace AJ,Jr, Bittner ML, Dougherty ER. 2005. Boolean relationships among genes responsive to ionizing radiation in the NCI 60 ACDS. Bioinformatics (Oxford, England) 21:1542-1549.

Pe'er D, Regev A, Elidan G, Friedman N. 2001. Inferring subnetworks from perturbed expression profiles. Bioinformatics (Oxford, England) 17 Suppl 1:S215-24.

Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM. 2007. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia (New York, N.Y.) 9:166-180.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129:1311-1323.

Roth RB, Hevezi P, Lee J, Willhite D, Lechner SM, Foster AC, Zlotnik A. 2006. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. Neurogenetics 7:67-80.

Sahoo D, Dill DL, Tibshirani R, Plevritis SK. 2007. Extracting binary signals from microarray time-course data. Nucleic acids research 35:3705-3712.

Schafer J and Strimmer K. 2005. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics (Oxford, England) 21:754-764.

Segal E, Friedman N, Kaminski N, Regev A, Koller D. 2005. From signatures to models: understanding cancer using microarrays. Nature genetics 37 Suppl:S38-45.

Segal E, Friedman N, Koller D, Regev A. 2004. A module map showing conditional activity of expression modules in cancer. Nature genetics 36:1090-1098.

Segal E, Taskar B, Gasch A, Friedman N, Koller D. 2001. Rich probabilistic models for gene expression. Bioinformatics 17:S243-252.

Sharief FS, Mohler JL, Sharief Y, Li SS. 1994. Expression of human prostatic acid phosphatase and prostate specific antigen genes in neoplastic and benign tissues. Biochemistry and molecular biology international 33:567-574.

Shmulevich I and Kauffman SA. 2004. Activities and sensitivities in boolean network models. Physical Review Letters 93:048701.

Shmulevich I and Zhang W. 2002. Binary analysis and optimization-based normalization of gene expression data. Bioinformatics (Oxford, England) 18:555-565.

Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED. 2004. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC bioinformatics 5:129.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular biology of the cell 9:3273-3297.

Stamenkovic I and Seed B. 1988. CD19, the earliest differentiation antigen of the B cell lineage, bears three extracellular immunoglobulin-like domains and an Epstein-Barr virus-related cytoplasmic tail. The Journal of experimental medicine 168:1205-1210.

Storey JD and Tibshirani R. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100:9440-9445.

Strand AD, Aragaki AK, Baquet ZC, Hodges A, Cunningham P, Holmans P, Jones KR, Jones L, Kooperberg C, Olson JM. 2007. Conservation of regional gene expression in mouse and human brain. PLoS genetics 3:e59.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science (New York, N.Y.) 302:249-255.

Swain M, Hunniford T, Dubitzky W, Mandel J, Palfreyman N. 2005. Reverse-engineering gene-regulatory networks using evolutionary algorithms and grid computing. Journal of clinical monitoring and computing 19:329-337.

Tamada Y, Bannai H, Imoto S, Katayama T, Kanehisa M, Miyano S. 2005. Utilizing evolutionary information and gene expression data for estimating gene networks with bayesian network models. Journal of bioinformatics and computational biology 3:1295-1313.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 1999. Systematic determination of genetic network architecture. Nature genetics 22:281-285.

Tirosh I, Weinberger A, Carmi M, Barkai N. 2006. A genetic signature of interspecies variations in gene expression. Nature genetics 38:830-834.

Tsaparas P, Marino-Ramirez L, Bodenreider O, Koonin EV, Jordan IK. 2006. Global similarity and local divergence in human and mouse gene co-expression networks. BMC evolutionary biology 6:70.

van Noort V, Snel B, Huynen MA. 2003. Predicting gene function by conserved co-expression. Trends in genetics : TIG 19:238-242.

Wang K, Nemenman I, Banerjee N, Margolin A, Califano A. 2005. Genome-wide discovery of modulators of transcriptional interactions in human B lymphocytes.

Weller PA, Critcher R, Goodfellow PN, German J, Ellis NA. 1995. The human Y chromosome homologue of XG: transcription of a naturally truncated gene. Human molecular genetics 4:859-868.

Figure legends

Figure 1. Boolean relationships: Six different types of Boolean relationships between pairs of genes taken from the Affymetrix U133 Plus 2.0 human dataset. Each point in the scatter plot corresponds to a microarray experiment, where the value for the x-axis is gene expression for the x-axis gene and the value for the y-axis is gene expression for the y-axis gene. There are 4,787 points in each scatter plot. (a) Equivalent relationship between CCNB2 and BUB1B. (b) PTPRC low ⇒ CD19 low. (c) XIST high ⇒ RPS4Y1 low. (d) Opposite relationship between EED and XTP7. (e) FAM60A low ⇒ NUAK1 high. (f) COL3A1 high ⇒ SPARC high.

Figure 2. Comparison of Boolean network with correlation-based network: On human CD (clusters of differentiation) genes: this plot shows the histogram of different types of Boolean relationships. (a) Equivalent. (e) Opposite. (b) Low ⇒ Low. (c) High ⇒ Low. (f) Low ⇒ High. (g) High ⇒ High. (d) No relationships. Example scatter plots of gene pairs with their correlation coefficient. (h) COL3A1 low ⇒ COL1A1 low, correlation coefficient = 0.933. This is an example of a clear asymmetric relationship with very high correlation coefficient. (i) VPREB1 high ⇒ IGLL1 high, correlation coefficient = 0.7963. This is an example of a clear asymmetric relationship with moderate correlation coefficient. (j) TLR2 and ITGAM are equivalent, correlation coefficient = 0.7. This is an example of equivalent relationship with low correlation coefficient. (k) LAIR1 and WAS are equivalent, correlation coefficient = 0.5158, is an example of equivalent relationship with very low correlation coefficient.

Figure 3. Properties of Boolean network: Log-log plot of the histogram of the probesets with respect to their number of Boolean relationships. Human Boolean network: (a) total, (b) symmetric, (c) asymmetric Boolean relationships. Conserved human and mouse Boolean network: (d) total, (e) symmetric, (f) asymmetric Boolean relationships. Conserved human, mouse and fruit fly Boolean network: (g) total, (h) symmetric, (i) asymmetric Boolean relationships.

Figure 4. Highly conserved Boolean relationships: Orthologous CCNB2 and BUB1B equivalent relationships: (a) Bub1 vs CycB in fruit fly, (b) Bub1b vs Ccnb2 in mouse, (c) BUB1B vs CCNB2 in human. Orthologus BUB1B high ⇒ GABRB1 low: (d) Bub1 vs Lcch3 in fruit fly, (e) Bub1b vs Gabrb1 in mouse, (f) BUB1B vs GABRB1 in human. Orthologous E2F2 ⇒ PCNA high: (g) E2f vs mus209 in fruit fly, (h) E2f1 vs Pcna in mouse, (i) E2F2 vs PCNA in human.

Figure 5. Boolean relationships follow known biology: (a) Gender difference, XIST high ⇒ RPS4Y1 low, male is different from female. (b) Gender tissue specific, RPS4Y1 low ⇒ ACPP low, only males have prostates. (c) Tissue difference, ACPP high ⇒ GABRB1 low, prostate is different from brain. (d) Development, HOXD3 high ⇒

HOXA13 low, anterior is different from posterior. (e) Differentiation, KIT high ⇒ CD19 low, Differentiated B Cell is different from HSC. (f) Co-expression, CDC2 vs CCNB2.

Figure 6. Boolean analysis: The expression levels of each probeset are sorted and a step function is fitted (using StepMiner) to the sorted expression level w minimizes the square error between the original and the fitted values. A threshold *t* is chosen, where the step crosses the original data. The region between *t*-0.5 and *t*+0.5 is classified as "intermediate", the region below *t*-0.5 is classified as "low" and the region above *t*+0.5 is classified as "high". The examples show probesets for two genes CDH1 and CDC2. As can be seen, CDH1 has a sharp rise between 6 and 9 and the StepMiner algorithm was able to assign a threshold in this region. CDC2, however, is very linear, and the StepMiner algorithm assigns the threshold approximately in the middle of the line. A scatter plot is shown to illustrate the analysis. Each point in the scatter plot corresponds to a microarray experiment, where the value for the x-axis is CDC2 expression and the value for the y-axis is CDH1 expression. Boolean analysis is performed on a pair of probesets, which ignores all the points that lie in the intermediate region and analyzes the four quadrants of the scatter plot. Four asymmetric relationships (low ⇒ low, low ⇒ high, high ⇒ low, high ⇒ high) are discovered, each corresponds to exactly one sparse quadrant in the scatter plot and two symmetric relationships (equivalent and opposite) are discovered each corresponds to two diagonally opposite sparse quadrants.

Table legends

Table 1: Number (in millions) of Boolean relationships in human, mouse and fruit fly datasets. The human dataset has 1% symmetric (equivalence + opposite) and 99% asymmetric (low ⇒ low + low ⇒ high + high ⇒ low + high ⇒ high) relationships of the total Boolean relationships. The mouse dataset has 1.4% symmetric (equivalence + opposite) and 98.6% asymmetric (low ⇒ low + low ⇒ high + high ⇒ low + high ⇒ high) relationships of the total Boolean relationships. The fruit fly dataset has 12% symmetric (equivalence + opposite) and 88% asymmetric (low ⇒ low + low ⇒ high + high ⇒ low + high ⇒ high) relationships of the total Boolean relationships.

Figures

Figure 1. Boolean relationships

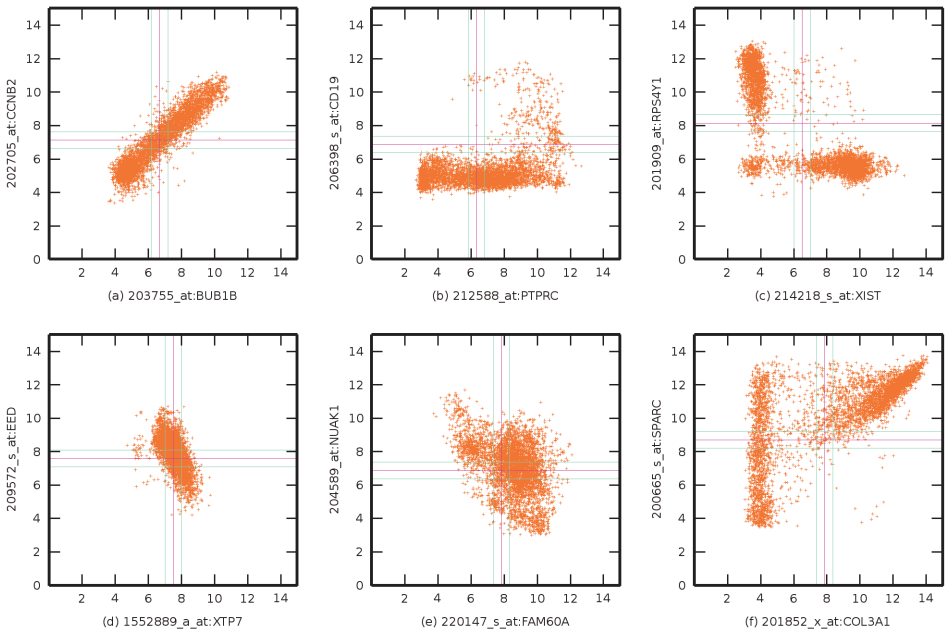


Figure 2. Comparison of Boolean network with correlation-based network

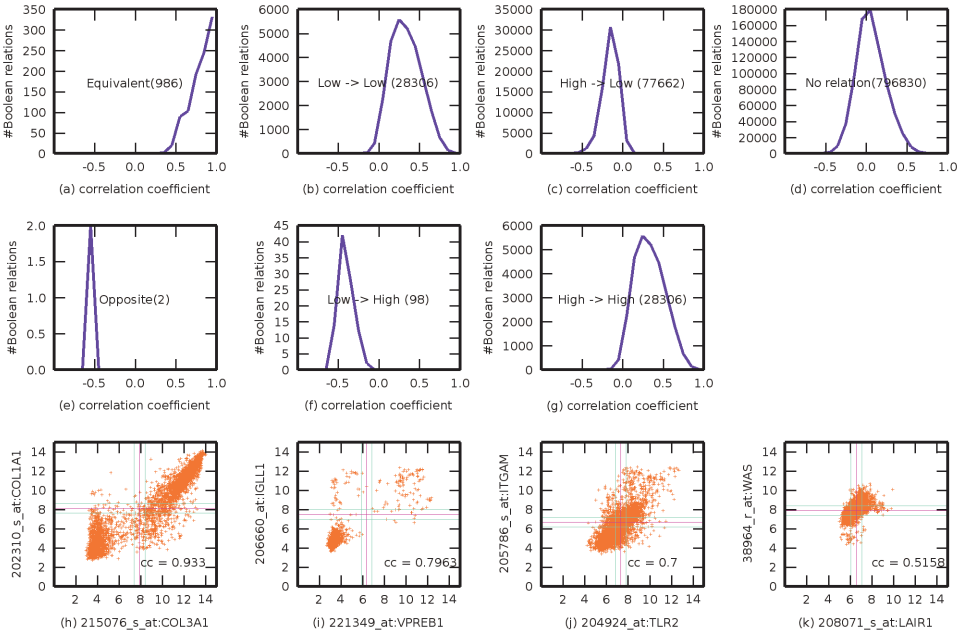


Figure 3. Properties of Boolean network

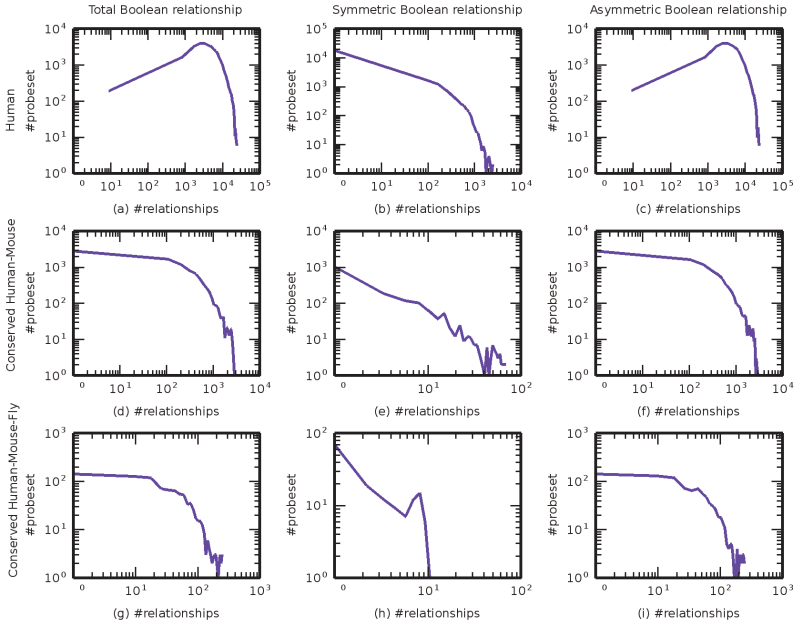


Figure 4. Highly conserved Boolean relationships

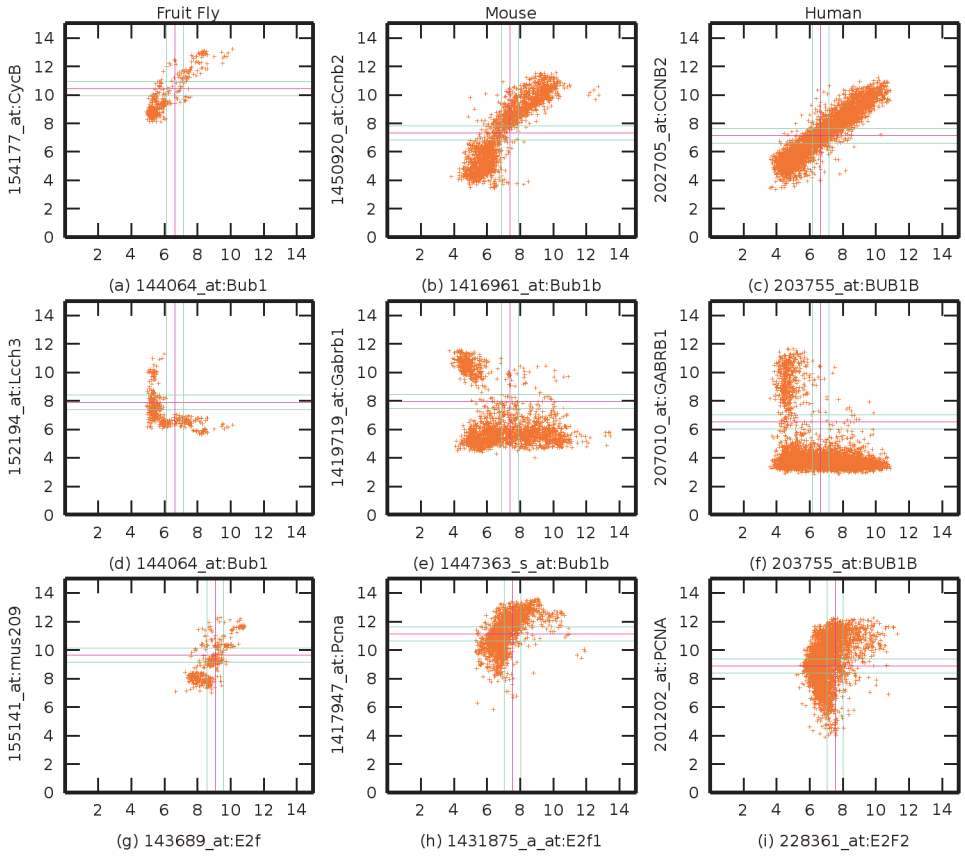


Figure 5. Boolean relationship follows known biology

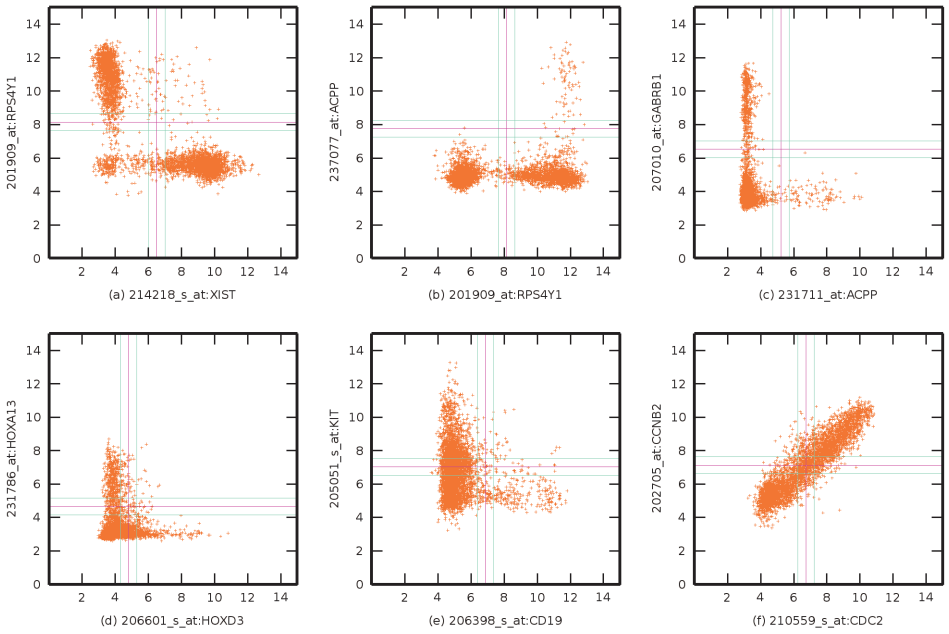
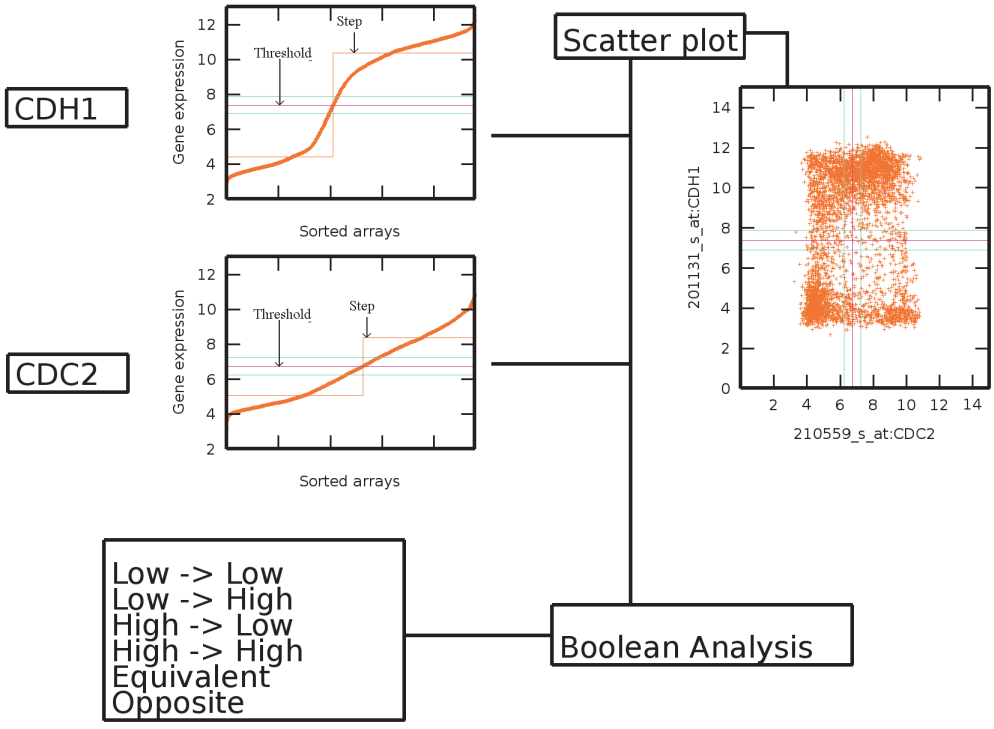


Figure 6. Boolean analysis



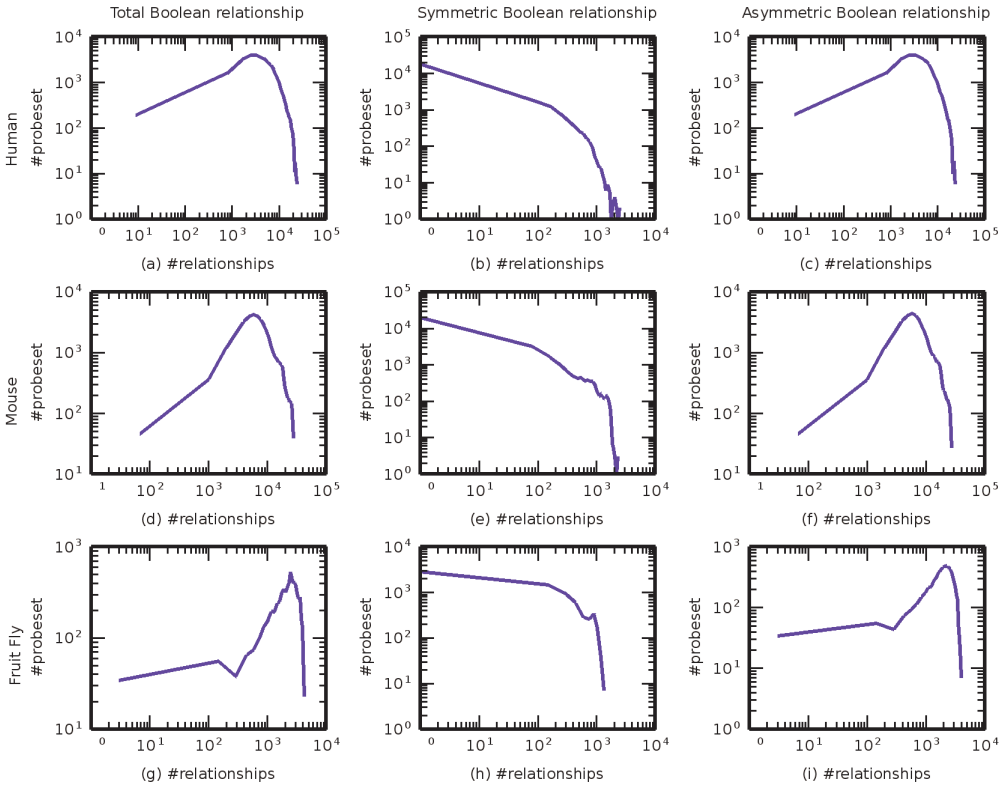
Tables

Table 1: Number (in millions) of Boolean relationships in human, mouse and fruit fly datasets.

Dataset	Total	Low implies High	High implies Low	Low implies Low	High implies High	Equivalent	Opposite
Human	208	2	128	38	38	1.6	0.4
Mouse	336	8	208	57.6	57.6	4.1	0.7
Fruit Fly	17	0.3	7.3	3.7	3.7	1.9	0.1

Supplementary information

Figure 1. Properties of human mouse and fruit fly Boolean networks: log-log plot of the histogram of the probesets with respect to their number of Boolean relationships. Human Boolean network: (a) Total, (b) symmetric, (c) asymmetric Boolean relationships. Mouse Boolean network: (d) Total, (e) symmetric, (f) asymmetric Boolean relationships. Fruit fly Boolean network: (g) Total, (h) symmetric, (i) asymmetric Boolean relationships.



Following files can be accessed at http://gourd.stanford.edu/~sahoo/recomb07/.

File 1. Connected component analysis: The cluster of genes can be found in each line as tab separated HUGO gene symbol name.

File 2. DAVID functional annotation (GO Analysis) on the largest cluster
File 3. DAVID functional annotation (GO Analysis) on the second largest cluster
File 4. DAVID functional annotation (KEGG) on the largest cluster
File 5. DAVID functional annotation (KEGG) on the second largest cluster