Global Similarity and Local Variance in Human Gene Coexpression Networks

Ivan Krivosheev[1], Lei Du[2], Hongzhi Wang[1], Shaojun Zhang[1,2], Yadong Wang[1,*], and Xia Li[1,2,*]

[1] School of Computer Science and Technology, Harbin Institute of Technology Harbin, 150001, China
[2] College of Bioinformatics Science and Technology, Harbin Medical University Harbin, 150081, China

Abstract. For the study presented here, we performed a comparative analysis of whole-genome gene expression variation in 210 unrelated HapMap individuals to assess the extent of expression divergence between 4 human populations and to explore the connection between the variation of gene expression and function. We used the GEO series GSE6536 to compare changes in expression of 47,294 human transcripts between four human populations. Gene expression patterns were resolved into gene coexpression networks and the topological properties of these networks were compared. The interrogation of coexpression networks allows for the use of a well-developed set of analytical and conceptual tools and provides an opportunity for the simultaneous comparison of variation at different levels of systemic organization, i.e., global vs. local network properties. The results of this comparison indicate that human co-expression networks are indistinguishable in terms of their global properties but show divergence at the local level.

Keywords: gene coexpression network interrogation.

1 Introduction

In the last few years, gene coexpression networks attracts attention of many researches[1,2]. According to previous studies, these networks are considered as a graph where each node represents a gene, and edges represent statistically high relationship between genes. The interaction between two genes in a gene network does not necessarily imply a physical interaction[4]. For the study presented here, we performed a comparative analysis of whole-genome gene expression variation in 210 unrelated HapMap individuals[7] to assess the extent of expression divergence between 4 human populations and to explore the connection between the variation of gene expression and function. Gene coexpression network could be constructed by using the GeneChip expression profiles data in NCBI GEO (Gene Expression Omnibus repository, database of gene expression data). [3]

* Corresponding authors.

H. Deng et al. (Eds.): AICI 2009, LNAI 5855, pp. 159–166, 2009.

2 Materials and Methods

2.1 Data Sets

Expression profiles of human gene pairs were compared in order to evaluate the divergence of human gene expression patterns. A total of 47,294 human transcripts for every population were considered. All-against-all gene expression profile comparisons for human populations' matrices (47294*60 CEU, 47294*45 CHB, 47294*45 JPT, and 47294*60 YRI) were used to generate population-specific coexpression networks. For coexpression networks, nodes correspond to genes, and edges link two genes from the same population if their expression profiles are considered sufficiently similar.

2.2 Network Construction

There are a number of existing reverse-engineering methods to construct coexpression network such as Relevance Networks, Bayesian networks etc. Results reported here are for networks constructed using Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE)[8], a novel information-theoretic algorithm for the reverse-engineering of transcriptional networks from microarray data. ARACNE infers interactions based on mutual information between genes, an information-theoretic measure of pairwise correlation. ARACNE compares favorably with existing methods and scales successfully to large network sizes.

ARACNE relies on a two-step process. First, candidate interactions are identified by estimating pairwise gene-gene mutual information (MI) and by filtering them using an appropriate threshold, I0, computed for a specific p-value, p0, in the null-hypothesis of two independent genes. This step is almost equivalent to the Relevance Networks method. Thus, in its second step, ARACNE removes the vast majority of indirect candidate interactions using a well-known property of mutual information – the data processing inequality (DPI) The DPI states that if genes g1 and g3 interact only through a third gene, g2, (i.e., if the interaction network is $g1<->\ldots<->g2<->\ldots<->g3$ and no alternative path exists between g1 and g3), then the following holds $I(g1,g3) \leq \min[I(g1,g2); I(g2,g3)]$.

For presented study we set mutual information threshold value at 0.001 and DPI value at 0.1. If the P value from this test is smaller than a predefined threshold, the edge will be preserved in the consensus network. When those threshold and DPI values are used, coexpression networks tend to congeal into graphs that are so densely connected as to preclude meaningful analysis of their topological properties.

3 Results and Discussion

3.1 Human Population Networks

Human population coexpression networks were evaluated with respect to a number of parameters describing their global topological properties and found to be highly similar (Table 1). For each network were computed a variety of parameters

provided by Cytoscape (open source bioinformatics software platform) plugin, Network Analyzer[9]. The numbers of nodes and edges in each network are comparable, with the CEU population network showing higher values for both. The average degree (<*k*>) is the average number of edges per node and gives rough approximation of how dense the network is. The CEU population network shows a slightly higher <*k*> which is consistent with the greater numbers of nodes and edges. However, <*k*> is again similar for all networks and rather high. By way of comparison, typical world-wide web networks have <*k*>≈7. The values of <*k*> might not be particularly relevant because, as will be shown below, the degree distributions are highly skewed.

A more refined notion of network density is given by the average clustering coefficient (<*C*>). The clustering coefficient *C* of a node *i* is defined as the fraction of the pairs of neighbors of node *i* that are linked to each other: $Ci = 2ni/ ki(ki-1)$, where *ni* is the number of observed links connecting the *ki* neighbors of node *i* and $ki(ki-1)/2$ is the total number of possible links. The average clustering coefficient (<*C*>) is the mean of this value for all nodes with at least two neighbors, and for all human population networks <*C*>≈0.3 (Table 1). For networks of this size, these <*C*> values are considered to be quite high. The high density of the coexpression networks is not necessarily surprising because, as one could reasonably expect, co-expression is, largely (but not entirely), transitive. In other words, if gene *A* is coexpressed with genes *B* and *C*, then genes *B* and *C* are likely to be coexpressed as well.

Table 1. Network parameters

Parameter	Population			
	CEU	CHB	JPT	YRI
Clustering coefficient	0.350	0.272	0.304	0.320
Network diameter	16	19	24	26
Network centralization	0.047	0.032	0.066	0.042
Average degree	4.576	10.721	16.46	13.781
Number of nodes	5546	3180	3572	3061
Number of edges	72073	17047	29398	21092
Network density	0.005	0.003	0.005	0.005
Network heterogeneity	1.539	1.601	1.778	1.622
Characteristic path length	25.991	6.297	6.733	7.246

3.2 Intersection Network

As described above, the human population gene coexpression networks are closely similar in terms of their global topological characteristics; they share similar node degree (k) distributions and $C(k)$ distributions as well as similar average node degrees ($\langle k \rangle$), clustering coefficients ($\langle C \rangle$) and path lengths ($\langle l \rangle$). Other parameters related to neighborhood, such as network density, network centralization and network heterogeneity are closely similar.

We further sought to evaluate the similarity between the population-specific coexpression networks at a local level. There is as yet no general method for assessing local network similarity (or graph isomorphism). However, in the case of the human population gene coexpression networks generated here, the use of orthologous gene pairs results in a one-to-one mapping between the nodes of the two networks. In this sense, the networks can be considered to be defined over the same set of nodes N, and thus can be directly compared by generating an intersection network. The human population intersection network is defined as the network over the set of nodes N where there is a link between two nodes i and j if i and j denote two pairs of orthologous genes which are connected in every human population network. Thus, the intersection network captures the coexpressed gene pairs conserved between 4 human populations.

The global characteristics of the intersection network are shown in Table 2. The intersection network node degree and $C(k)$ distributions are clearly similar to those of the population-specific networks as are the average clustering coefficient ($\langle C \rangle$ = 0.213) and average path length ($\langle l \rangle$ = 3.04). Network diameter equals 10. The network diameter and the average shortest path length, also known as the characteristic path length, indicate small-world properties of the analyzed network. Taken together, these findings indicate that the global structure of the population-specific coexpression networks is preserved in the intersection network. However, the most striking feature of the intersection network is the small fraction of genes (~20%) and edges (~4–16%) that are conserved between populations networks (Table 3). Accordingly, the average node degree is lower ($\langle k \rangle$ = 7.518) in the intersection network than it is in each of the population-specific networks.

Table 2. Number of nodes and edges in intersection network

	Nodes	Edges
Intersection network	713	2680
CEU	5546 (13%)	72073 (4%)
CHB	3180 (22%)	17047 (16%)
JPT	3572 (20%)	29398 (9%)
YRI	3061 (23%)	21092 (13%)

Table 3. Intersection network parameters

Parameter	Intersection network	Random network
Clustering coefficient	0.213	0.001
Network diameter	10	10
Network centralization	0.061	-
Average degree	7.518	0.0
Number of nodes	713	713
Number of edges	2680	2680
Network density	0.011	-
Network heterogeneity	1.592	-
Characteristic path length	3.040	0.005

Random network generated according Erdos-Renyi model, it has same number of nodes and edges with intersection network.

3.3 Functional Coherence of Gene Coexpression Networks

Genes in the networks were functionally categorized using their Gene Ontology (GO) biological process annotation terms. Overrepresented GO terms were identified with BINGO[14] by comparing the relative frequencies of GO terms in specific clusters with the frequencies of randomly selected GO-terms. The Hypergeometric test was used to

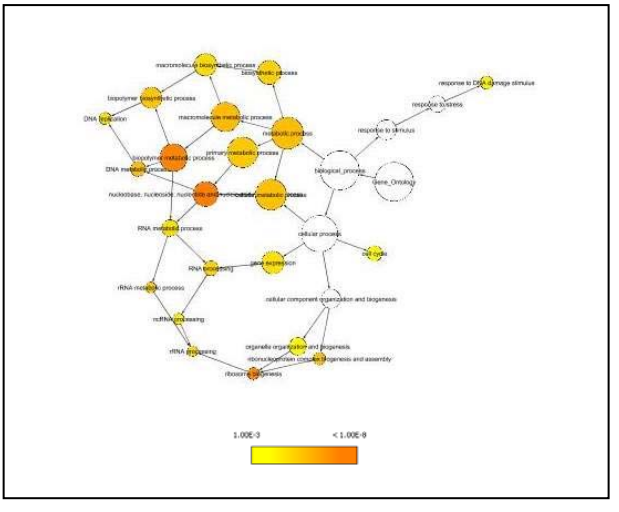

Fig. 1. Overrepresented GO terms

do this with the Benjamini and Hochberg false discovery rate correction for multiple tests and a *P*-value threshold of 0.001. Pairwise similarities between gene GO terms were measured using the semantic similarity method, which computes the relative distance between any two terms along the GO-graph . Result is shown in Table 4.

The graph (Figure 1) visualizes the GO categories that were found significantly over-represented in the context of the GO hierarchy. The size (area) of the nodes is proportional to the number of genes in the test set which are annotated to that node. The color of the node represents the (corrected) *p*-value. White nodes are not significantly over-represented, the other ones are, with a color scale ranging from yellow (*p*-value = significance level, e.g. 0.001) to dark orange (*p*-value = 5 orders of magnitude smaller than significance level, e.g. 10^{-5} * 0.001). The color saturates at dark orange for *p*-values which are more than 5 orders of magnitude smaller than the chosen significance level.

Table 4. GO process annotation

GO-ID	Description	p-value
43283	biopolymer metabolic process	3.4840E-11
42254	ribosome biogenesis	8.2487E-11
43170	macromolecule metabolic process	7.3685E-9
6259	DNA metabolic process	1.1960E-8
8152	metabolic process	1.3627E-8
44237	cellular metabolic process	2.3559E-8
22613	ribonucleoprotein complex biogenesis and assembly	4.3938E-8
43284	biopolymer biosynthetic process	4.6317E-8
16072	rRNA metabolic process	6.9420E-8
44238	primary metabolic process	7.6631E-8
9058	biosynthetic process	1.9158E-7
6394	RNA processing	1.9444E-7
6365	rRNA processing	4.0186E-7
16070	RNA metabolic process	6.0819E-7
9059	macromolecule biosynthetic process	6.2404E-7
10467	gene expression	9.9307E-7
6260	DNA replication	1.1083E-6
34470	ncRNA processing	3.6057E-6
6974	response to DNA damage stimulus	6.6203E-6
7049	cell cycle	1.0995E-5
6996	organelle organization and biogenesis	1.4177E-5

In fact, from the figure it could be seen that the category ‘biopolymer metabolic process’ is the important one, and that the over-representation of ‘macromolecule metabolic process ' and 'metabolic process ' categories is merely a result of the presence of those 'protein modification' genes. The fact that both categories are colored equally dark, is due to the saturation of the node color for very low *p*-values.

4 Conclusion

The global topological properties of the human population gene coexpression networks studied here are very similar but the specific architectures that underlie these properties are drastically different. The actual pairs of orthologous genes that are found to be co-expressed in the different population are highly divergent, although we did detect a substantial conserved component of the co-expression network. One of the most prevalent functional classes that show clear function-expression coherence are genes involved in biopolymer metabolism. Example of these cluster is shown in Figure 2.

The biological relevance of the global network topological properties appears questionable[10]. Of course, this does not prevent network analysis from being a powerful approach, possibly, the most appropriate one for the quantitative study of complex systems made up of numerous interacting parts.

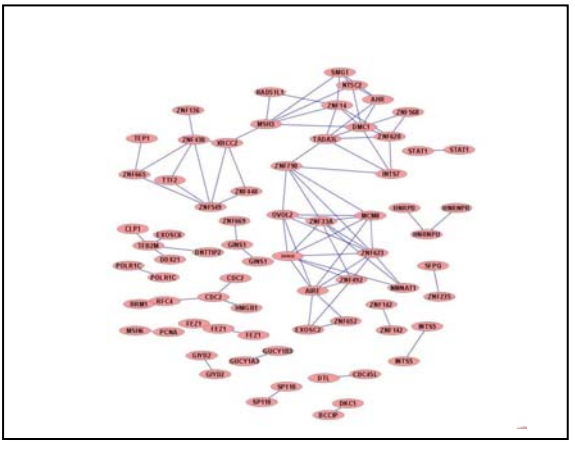

Fig. 2. Cluster of tightly coexpressed and functionally coherent genes

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 30871394, 30370798 and 30571034), the National High Tech Development Project of China, the 863 Program (Grant Nos. 2007AA02Z329), the National Basic Research Program of China, the 973 Program (Grant Nos. 2008CB517302) and the National Science Foundation of Heilongjiang Province (Grant Nos. ZJG0501, 1055HG009, GB03C602-4, BMFH060044, and D200650).

References

1. Horvath, S., Dong, J.: Geometric interpretation of gene coexpression network analysis. PLoS Comput. Biol. 4(8), e1000117 (2008)
2. Carter, S.L., et al.: Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 20(14), 2242–2250 (2004)
3. Bansal, M., et al.: How to infer gene networks from expression profiles. Mol. Syst. Biol. 3, 78 (2007)
4. Potapov, A.P., et al.: Topology of mammalian transcription networks. Genome Inform. 16(2), 270–278 (2005)
5. Yu, H., et al.: The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput. Biol. 3(4), e59 (2007)
6. Stranger, B.E., et al.: Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813), 848–853 (2007)
7. The International HapMap Project. Nature 426(6968), 789–796 (2003)
8. Margolin, A.A., et al.: ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7(suppl. 1), S7 (2006)
9. Vlasblom, J., et al.: GenePro: a Cytoscape plug-in for advanced visualization and analysis of interaction networks. Bioinformatics 22(17), 2178–2179 (2006)
10. Tsaparas, P., et al.: Global similarity and local divergence in human and mouse gene co-expression networks. BMC Evol. Biol. 6, 70 (2006)
11. Khaitovich, P., et al.: A neutral model of transcriptome evolution. PLoS Biol. 2(5), E132 (2004)
12. Yanai, I., Graur, D., Ophir, R.: Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS 8(1), 15–24 (2004)
13. Jordan, I.K., Marino-Ramirez, L., Koonin, E.V.: Evolutionary significance of gene expression divergence. Gene 345(1), 119–126 (2005)
14. Babu, M.M.: Introduction to microarray data analysis. In: Grant, R.P. (ed.) Computational Genomics: Theory and Application. Horizon Press, Norwich (2004)