Extraction of Developmentally Important Genes from Microarray Data

Dong-Soo Kahng[1*], Tae Woo Ryu[1*], A Reum Han[1*], Hyun S. Moon[2], Kwang H. Lee[1] and Doheon Lee[1§]
[1]Department of BioSystems and [2]Department of Computer Science
Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea
[*]These authors contributed equally to this work.
[§]To whom correspondence should be addressed.
email:dskahng@bioif.kaist.ac.kr; doheon@kaist.ac.kr

***Abstract*—Using microarray data of 4,028 genes of *Drosophila melanogaster* during the life-cycle, we constructed gene expression networks for four developmental stages of the fruitfly: egg-early embryo, embryo, larva and pupa. The network for each stage showed a scale-free property with 0.85 < γ < 1.85 and revealed one or two giant clusters and many small clusters. Since the hubs are thought to be important in the network, we analyzed genes with high degree, hubs, for all network and found many previously studied genes that have specific functions in each stage. We also assigned the biological process of gene ontology (GO) to neighbors of hubs and found that many clusters have stage-specific characteristics.**

I. Introduction

The development of an animal proceeds from the spatiotemporal expression of many genes. Elucidation of the overall process is necessary to determine targets for the medical treatment of many genetic diseases, as well as for the comprehensive understanding of development that many genes are involved in. In spite of its importance, it is difficult to find the global pattern of development of an animal because previous researches, especially in molecular biology, have concentrated on an individual gene and a small number of interacting proteins.

Recently, the expression pattern for 4,028 genes during the life cycle of *Drosophila melanogaster* was studied using microarray experiments [2]. Microarray is technology for monitoring abundance of gene expression in cells. This was the first study to provide the global gene expression pattern of a higher organism during its whole life. Using this data, the gene networks of each development stage were constructed and analyzed with respect to the characteristics of the network itself.

Many complex networks in nature have properties of scale-free networks [1]. Biological networks, including protein-protein interaction [10], orthologue conservation and the metabolic pathway, are also known as scale-free networks [1]. A scale-free network has a degree distribution P(k) which is proportional to the –γ th power of k, that is, a power-law distribution (Degree means the number of links which a node has in a network, and k denotes degree).

$$P(k) \sim k^{-\gamma} \tag{1}$$

Those networks are composed of many nodes of small degree and a few nodes of large degree. The latter are known as hubs, which are thought to play an important role in the network. Our networks also showed scale-free properties, which meant that we might be able to define the hub genes for each stage.

II. Methods

A. Data preparation

We acquired microarray time series data of *D. melanogaster* [2], and divided the data into four stages: egg-embryo, embryo, larva, and pupa. The egg-embryo stage data were composed of one sample from the egg and 10 samples from the early embryo. In addition, we added whole data to compare with each stage, to make five datasets in all. Missing values were filled with the average value of the gene expression in each stage.

B. Construction of Network from data

For the microarray data of *D.melanogaster*, we assigned each gene to a node for a network, and calculated a score, ρ_{XY}, of each pair of gene X and Y for 4,028 genes. The score is calculated using Pearson correlation coefficient:

$$\frac{\sum_{i}^{N} X_i Y_i - \frac{\sum_{i}^{N} X_i \sum_{i}^{N} Y_i}{N}}{\sqrt{\left(\sum_{i}^{N} X_i^2 - \frac{(\sum_{i}^{N} X_i)^2}{N}\right)\left(\sum_{i}^{N} Y_i^2 - \frac{(\sum_{i}^{N} Y_i)^2}{N}\right)}} \qquad (2)$$

(N: Number of sampling time,

Xi: i th observed value of X gene expression series,

Yi: i th observed value of Y gene expression series)

Then, we sorted the scores and connected two nodes which have the highest score one by one until the number of clusters becomes a maximum [15].

C. Extraction and interpretation of hubs and strong neighbors

We defined the hub gene, which is in the top 1% with high degree, from each network. Neighbors of a hub gene are genes that have a path length of 1 to the hub. We also defined a strong neighbor, which is a neighbor of hub and connected at least one other neighbor (See Fig.1).

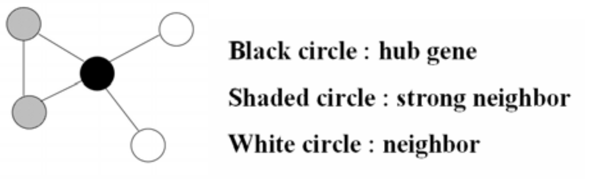

Figure 1. Hub and strong neighbors

To analyze hubs and neighbors, we got the gene annotation information from flybase [17] and Gene Ontology database [18].

III. RESULT

A. Whole and Stage specific Networks

We divided the microarray gene expression data of the life cycle of *D. melanogaster* into five datasets, comprising four life stages and the whole life cycle and construct a network for each stage. Most networks for each stage consisted of one or two big clusters which have several hundreds nodes, and several tens or hundreds clusters which had fewer than ten nodes. We analyzed big clusters, extracted sub-clusters by connectivity density, and observed the degree distribution for each node. In the ln k vs. ln P(k) graph (See Fig. 2), most clusters showed a linear relation with minus tangent. Such linear relation of the graph demonstrates the scale-free property of the networks [1].

For $P(k) \sim k^{-\gamma}$, $0.85 < \gamma < 1.85$ for all our cases. R^2 is the coefficient of determination ($0 \leq R^2 \leq 1$). A large value of R^2 tends to indicate that the data points are closer to the regression line [8].

B. Analysis of hubs and strong neighbors

For the networks of the five datasets, we defined hub genes, and assigned a function using Gene Ontology database (see method for detail). These are listed in table 1. In a number of stages, we found previously studied genes that have important roles in each stage.

In the hub list of the egg-embryonic stage, *prospero* (*pros*) and *tartan* (*trn*) have been known to affect the early embryonic development of the nervous system [3], [5]. *Tailup* (*tup*) is known to mediate the torso receptor pathway in the terminal region of the embryo [16]. *Wingless* (*wg*) serves as the major signaling molecule in embryonic patterning [19]. *18-wheeler* (*18w*) participates in segmentation in the embryo [6]. *Arrest* (*aret*) affects posterior body patterning by inhibiting the translation of *oskar* mRNA [11]. At the embryonic stage, *Rpd3* has a role in embryonic pattern formation [4]. *Caf1* is a member of the Polycomb group and by inhibiting the transcription of the homeotic gene they play a role in early embryonic patterning [13]. *Pp4-19C* is required for regulation of the cell cycle in the embryo [9]. At the pupal stage, *inaF*, eye-enriched protein, plays a role in the rhodopsin-mediated signaling [12]. *Mhc* is known to play a role in muscle fiber differentiation in the pupal stage [7].

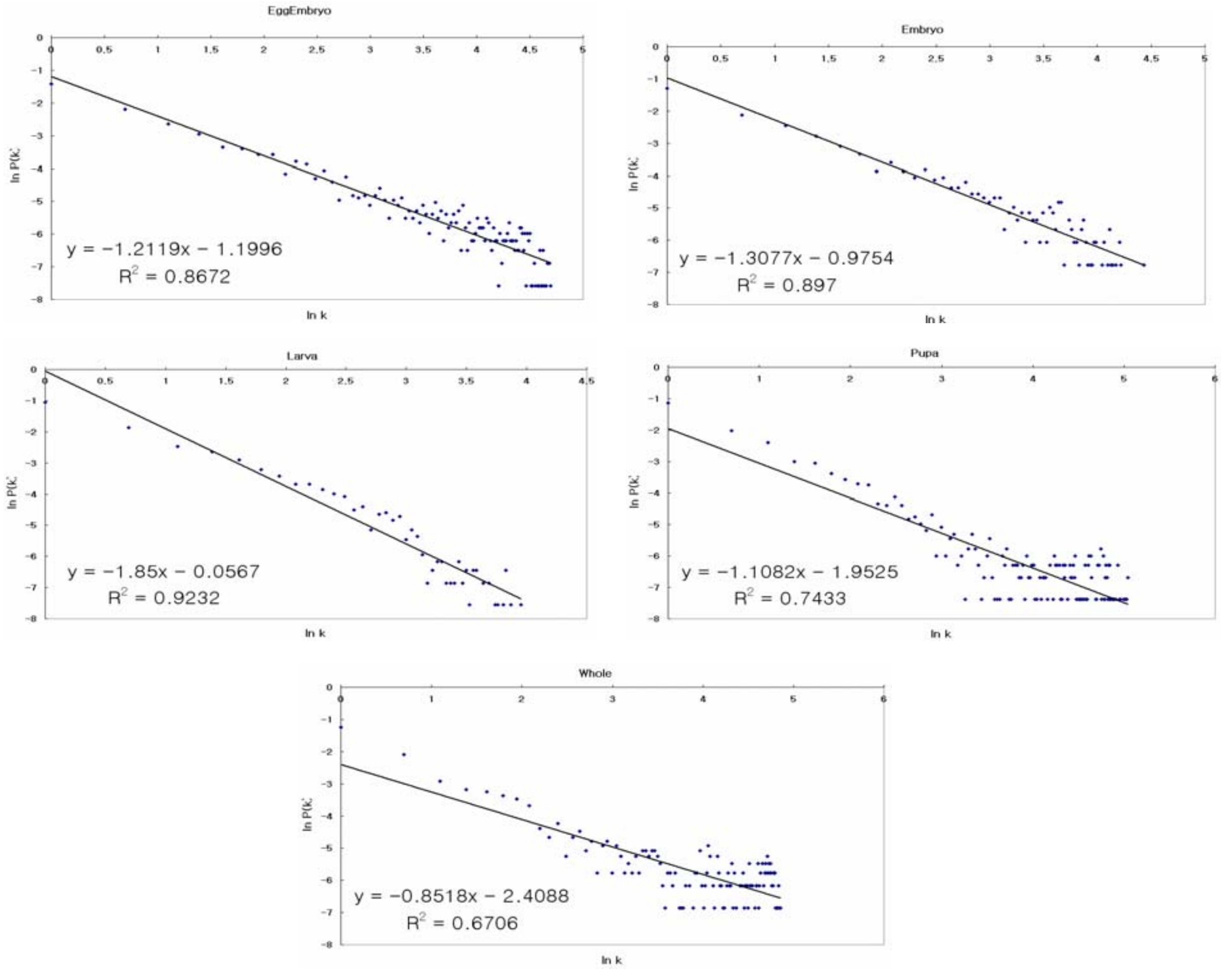

Figure 2. Scale-free property of networks. From upper left to lower right, they represent egg-embryo, embryo, larva and pupa stages, and whole life, respectively. The x axis is ln k, and the y axis is ln P(k). The tangents represent γ: 0.85 < γ < 1.85. A large value of R^2 tends to indicate that the data points are closer to the regression line.

Table 1. Hubs and their Functions.

Annotated Biological Process is assigned using the Gene Ontology. See text for details

Gene Name	# of degree	Annotated Biological Process
Stage : Egg-Embryo		
RpL22	109	protein biosynthesis
CG5844	108	unknown
pros	108	axonogenesis, central nervous system development, dendrite morphogenesis, glial cell differentiation, peripheral nervous system development, regulation of neuron differentiation
mRpS31	107	protein biosynthesis
CG17506	107	nucleobase, nucleoside, nucleotide and nucleic acid metabolism, transcription, RNA-dependent
CG6287	105	unknown
CG2789	103	unknown
CG9578	102	unknown
trn	102	cell migration, tracheal system development (sensu Insecta)
CG13096	102	protein biosynthesis, protein metabolism
tup	101	terminal region determination, torso receptor signaling pathway
CG2145	100	unknown
CG15658	99	cell adhesion, transmission of nerve impulse
CG5002	99	unknown
wg	99	frizzled-2 receptor signaling pathway

(Table 1. Continued.) **Stage :** Egg-Embryo		
CG9246	98	unknown
18w	97	cell adhesion
aret	95	negative regulation of translation, oogenesis (sensu Insecta), spermatid development
CCR4	94	mRNA catabolism, deadenylation-dependent, regulation of transcription from Pol II promoter
Eno	94	glycolysis
Stage : Embryo		
Rpd3	84	chromatin silencing histone methylation
Caf1	68	negative regulation of transcription of homeotic gene (Polycomb group), chromatin silencing, histone acetylation, histone methylation, nucleosome mobilization, nucleosome spacing, transcription
CG4857	67	cell communication, signal transduction
Pp4-19C	67	M-phase specific microtubule process, microtubule-based process, regulation of mitotic cell cycle
cdc2	65	G2/M transition of mitotic cell cycle
Nup358	63	unknown
CG1078	62	nucleobase, nucleoside, nucleotide and nucleic acid metabolism, transcription from Pol II promoter
CG7357	61	nucleobase, nucleoside, nucleotide and nucleic acid metabolism, regulation of transcription from Pol II promoter, transcription from Pol II promoter
CG11982	61	unknown
FK506-bp1	61	protein folding
Stage : Larva		
l(2)35Di	52	unknown
CG17470	48	mesoderm development
CG9920	46	'de novo' protein folding
CG17838	46	unknown
CG12699	46	unknown
CG1324	45	cytoskeleton organization and biogenesis
CG8040	43	unknown
CG8701	42	unknown
CG6662	40	unknown
CG2149	40	cell communication, signal transduction
Stage : Pupa		
CG6439	155	tricarboxylic acid cycle
fln	155	unknown
inaF	153	maintenance of rhodopsin mediated signaling, rhodopsin mediated signaling
CG9090	152	phosphate transport
CG12233	150	tricarboxylic acid cycle
CG1826	148	unknown
Mhc	143	striated muscle contraction
CG9921	142	unknown
scpr-B	139	unknown
CG4975	139	unknown
CG10949	139	unknown
l(2)35Di	136	oxidative phosphorylation, mitochondrial electron transport, NADH to ubiquinone
CG8154	133	mesoderm development
CG9813	132	unknown
Cpn	131	unknown
Scp1	129	unknown
Stage : Whole		
CG2149	129	cell communication, signal transduction
CG8813	127	lipid metabolism, mRNA transcription, nucleobase, nucleoside, nucleotide and nucleic acid metabolism, steroid metabolism, transcription from Pol II promoter
CG5755	126	transport
CG7251	126	cytoskeleton organization and biogenesis, intracellular protein transport, protein metabolism, protein-mitochondrial targeting, proteolysis and peptidolysis
CG13030	124	unknown
Tsp3A	123	unknown
CG1394	122	unknown
CG9130	121	unknown
CG1340	121	translational initiation
CG5398	121	unknown

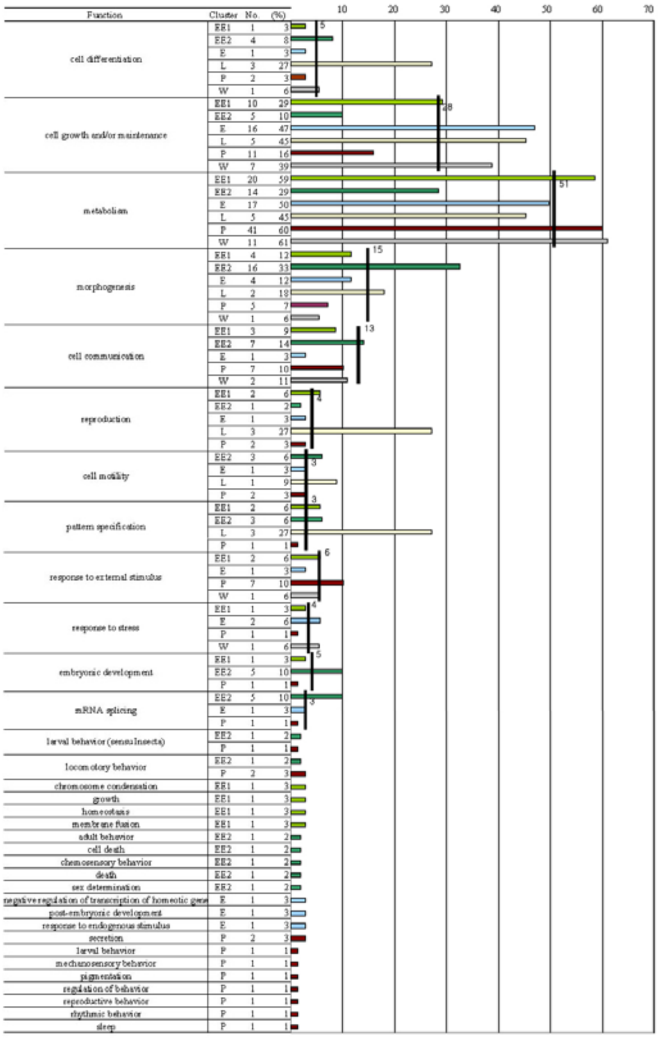

Figure 3. Functional classification of each cluster. EE1: egg-embryo cluster 1, EE2: egg-embryo cluster 2, E: embryo, L: larva, P: pupa, W: whole, No: number of genes in that cluster having that function. The 4th level biological process in GO is used to assign functions to genes. Percentages for each cluster are calculated considering only genes that have 4th level GO annotations. The black vertical line indicates the percentage of 4th level GO of the total number of genes in the microarray.

By contrast, we did not identify any known genes in the networks of the larval stage and whole life. Therefore, our approach is thought to be advantageous for finding developmentally important genes that cannot be found by considering only the whole life.

We also examined strong neighbors of hubs to determine whether they have stage-specific characteristics. Strong neighbors of hubs in the same cluster are also have high degree (data not shown), which might be due to the preferential attachment property of scale-free networks [1]. So, we found clusters including hubs from table 1, and from those clusters we extracted strong neighbors of the largest hub. Two groups of strong neighbors were detected in only the egg-early embryo stage. We classified the function of the strong neighbor genes for all networks (Fig. 3). Strong neighbors in the embryo were enriched of cell growth and/or maintenance compared with those of all genes (47% versus 28%), and strong neighbors of egg-embryo cluster 2 had low percentages in cell growth and/or maintenance (10% versus 28%) and metabolism (29% versus 51%) but were high in morphogenesis (33% versus 15%). The six genes of 65 that were annotated as embryonic development were detected in the two egg-embryo clusters. The larva cluster also showed unique high percentages in cell differentiation, reproduction and patter specification. However, this result might be due to the small number of annotated genes. Although there are limitations to analysis using GO because of the existence of non-annotated genes and multiple categorized genes, we were able to establish that strong neighbors can also show stage-specific characteristics. Moreover, our analysis was more informative than mere examination of the data as a whole.

IV. DISCUSSION

To detect developmentally important genes well, it is necessary to develop a method for constructing networks. First, proper gene filtering is required to eliminate the effect of constantly highly or lowly expressed genes in calculation scores. Secondly, because the Pearson correlation cannot contain sequential information, local clustering scoring might be more adequate for this kind of time series data [14].

We used gene ontology for the analysis of hubs and strong neighbors. While gene ontology is developing as an area of research, it has many limitations, as described above. Hence, we mined the literature to interpret hubs and found specific functions of some genes that are not described in the GO. Advanced gene ontology for developmental biology would have great utility.

Our proposed method and results are thought to be helpful for finding developmentally important genes. Moreover, as may be seen from table 1, genes whose functions are not yet known would be strong candidates for important genes of fruitfly development.

ACKNOWLEDGEMENTS

We thank Chung Moon Soul, the generous founder of the Department of BioSystems, KAIST, and members of the Department of Genetics, Yale University for contributing useful public microarray gene expression data of the life cycle of *D. melanogaster.*

REFERENCES

[1] Albert R. and Barabási A.L. Statistical mechanics of complex networks. *Reviews of modern physics*, vol.74, pp. 47-98, 2002.

[2] Arbeitman M.N., Furlong E.E.M., Imam F., Johnson E., Null B.H., Baker B.S., Krasnow,M.A., Scott,M.P., Davis,R.W. and White,K.P. Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, vol.297**,** pp.2270-2275, 2002.

[3] Chang Z., Price B.D., Bockheim S., Boedigheimer M.J., Smith R. and Laughon A. Molecular and genetic characterization of the Drosophila tartan gene. *Dev. Biol.,* vol. 160(2), pp. 315-332, 1993.

[4] Chen G., Fernandez J., Mische S. and Courey A.J. A functional interaction between the histone deacetylase Rpd3 and the corepressor groucho in Drosophila development. *Genes Dev.*, vol.13(17), pp. 2218-2230, 1999.

[5] Demidenko Z., Badenhorst P., Jones T., Bi X. and Mortin MA. Regulated nuclear export of the homeodomain transcription factor Prospero. *Development,* vol.128(8), pp.1359-1367, 2001.

[6] Eldon E., Kooyer S., D'Evelyn D., Duman M., Lawinger P., Botas J. and Bellen H. The Drosophila 18 wheeler is required for morphogenesis and has striking similarities to Toll. *Development,* vol.120(4), pp.885-899, 1994.

[7] Generalova M.V., Kriukova M.E. and Miasniankina E.N. (1994) An electron microscopic study of the structure of the indirect flight musculature at the pupal stage in the muscle mutant of *Drosophila melanogaster*. *Ontogenez*, vol.25(6), pp.33-41, 1994.

[8] Hayter A. *Probability and Statistics for Engineers and Scientists,* 2nd. Ed., Duxbury, USA, 2002.

[9] Helps N.R., Brewis N.D., Lineruth K., Davis T., Kaiser K. and Cohen P.T. Protein phosphatase 4 is an essential enzyme required for organisation of microtubules at centrosomes in *Drosophila* embryos. *J. Cell Sci.,* vol.111(10), pp.1331-1340, 1998.

[10] Jeong H., Mason S.P., Barabási A.L. and Oltvai Z.N. Lethality and centrality in protein networks. *Nature*, vol.411, pp.41-42, 2001.

[11] Kim-Ha J., Kerr K. and Macdonald P.M. Translational regulation of oskar mRNA by bruno, an ovarian RNA-binding protein, is essential. *Cell*, vol.81(3), pp.403-412, 1995.

[12] Li C., Geng C., Leung H.T., Hong Y.S., Strong L.L., Schneuwly S. and Pak W.L. INAF, a protein required for transient receptor potential Ca(2+) channel function. *Proc. Natl. Acad. Sci. USA*, vol.96(23), pp.13474-13479, 1999.

[13] Muller J., Hart C.M., Francis N.J., Vargas M.L., Sengupta A., Wild B., Miller E.L., O'Connor M.B., Kingston R.E. and Simon J.A. Histone methyltransferase activity of a *Drosophila* Polycomb group repressor complex. *Cell,* vol.111(2), pp.197-208, 2002.

[14] Qian J., Dolled-Filhart M., Lin J., Yu H. and Gerstein M. Beyond Synexpression Relationships: Local Clustering of Time-shifted and inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. *J. Mol. Biol.,* vol.314, pp.51053-51066, 2001.

[15] Rho K., Jeong H. and Kahng B. Identification of essential and functionally modeled genes through the microarray. condmat/0301110, 2003.

[16] Strecker T.R., Yip M.L. and Lipshitz H.D. Zygotic genes that mediate torso receptor tyrosine kinase functions in the *Drosophila melanogaster* embryo. *Proc. Natl. Acad. Sci. USA,* vol.88, pp.5824-5828, 1991.

[17] The FlyBase Consortium , The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Research,* vol.31**,** pp.172-175, http://flybase.org/, 2003

[18] The Gene Ontology Consortium, Gene Ontology: tool for the unification of biology. *Nature Genet*., vol.25, pp.25-29, 2000.

[19] Wodarz A. and Nusse R. Mechanisms of Wnt signaling in development. *A. Rev. Cell Dev. Biol.,* vol.14, pp.59-88, 1998.