2009－BIO－16 (5)
2009／3／5

ミツバチ幼虫成長過程のたんぱく質発現の非計量多次元尺度構成法による解析

田口善弘

〒 112-8551 東京都文京区春日 1-13-27 中央大学理工学部物理学科

Abstract

非計量多次元尺度構成法 (nMDS) をミツバチの幼虫の成長過程におけるたんぱく質発現の解析に用いた。共発現しているたんぱく質のセットの選択がより適切であることが分かった。また、共発現ネットワークを描くことに成功した。また、個々のたんぱく質発現プロファイルの関係を可視化することにも成功した。

Analysis of protein expression during honey bee larval development via non-metric multidimensional scaling method

Y-h. Taguchi[1]

Department of Physics, Chuo University, Kasuga 1-13-27, Bunkyo-ku, Tokyo 112-8551, Japan.

Abstract

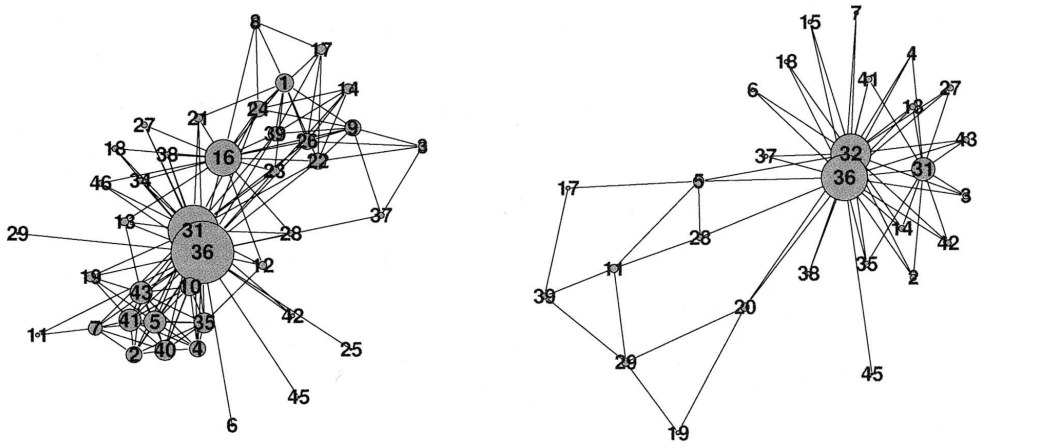
In this paper, we have applied non-metric multidimensional scaling method (nMDS) to protein expression profiles during bee larval development. We have found that we can select more reliable set of protein classes which are regarded as being co expressed. Applying the significance test for coexpressoion pairs, we have drawn coexpression network for these protein classes, which turns out to be very reasonable. nMDS also gives us visual representation of relationship between individual protein expressions, which helps us to understand protein expressions.

1 Introduction

Honey bee is an important insect for human beings due to applications, since it can produce honey as its names says. In addition to this, it is also interesting to study it as an example of social insects. Since the completeness of whole sequencing of honey bee genome[1], honey bee started to be investigated by exhaustive analysis. However, most of the researches are concentrated to adult insects[2, 3, 4, 5]. Recently, Chan and Foster[6] published the research of protein expressions during honey bee larval development via mass spectrometry-based measurements. Although this is the first investigation about larval, the expression analysis is hard due to lack of macroscopic morphological changes. In this paper, we have applied non-metric multidimensional scaling method (nMDS)[7] to figure out what have happened during larval development.

2 Results

Figures 1 shows the inter class coexpression network (see Methods and Materials). These classes are sets of annotated proteins by Chan and Foster[6]. They have manually curated proteins and given 47 protein classes. Since they are displayed via Fruchterman and Reingold method, tightly (i.e., via more edges) connected nodes are located loser to each other. First of all, for both of tissue and hemolymph, there are two hub protein classes, ribosome (ID 36) and proteasome (ID 31). We think that it is inevitable since their expression is protein based. No genes are measured if not translated. Thus, these two hub nodes do not express anything biologically significant. In contrast to this, electron transport (ID 16) is a hub only in Fig. 1(a) and protein folding (ID 32) is a hub in Fig. 1(b). This represents the difference between tissue samples and hemolymph. Tissue samples mainly contribute to energy production, while hemolymph samples do for protein synthesis. On the other hand, some of the proteins (e.g., amino acid metabolism (ID 6), fatty acid synthesis (ID 18), and nucleotide metabolism (ID 27)) are linked only to these hub nodes. Thus, these can be regarded as being isolated essentially. Most of them are related to metabolism, thus clearly metabolic network does not fully develop during larval development phases. Besides these isolated proteins, there seem to be some groups of nodes (proteins) connected with each other. For tissue samples (Fig.1(a)), there are two such groups. The group including



(a) (b)

Figure 1: Network representation of coexpressions of protein classes, (a) Tissue (b) Hemolymph. The numbers are ID listed in Tables 1 and 2. The radius of node is proportional to the number of edges linked to each node.

transcription (ID 41), Aldo-keto reductase superfamily (ID 5), ubiquitination (ID 43), tRNA synthetase (ID 40), and so on, which locates at the left-below region, is related to transcription-translation processes. On the other hand, the other group including ATP synthesis (ID 1), kinases or phosphatases (ID 24), mevalonate pathway (ID 26) and so on, which locates at the upper-right region, is related to energy production. For hemolymph samples, there is a group including mevalonate pathway (ID 26), carbohydrate metabolism (ID 11), glycolipid metabolism (ID 20) and so on, which locate lower-left, is again related to energy production. There are no groups related to transcription-translation related groups. Thus, our method can capture basic inter-protein class relationship well.

3 Discussion

The co expression significance within each protein class (intra protein co expression) is computed based upon embedings by nMDS (see Materials and Methods) and is shown in Tables 1 and 2. In Table 1, there are 7 protein classes which are significant due to all of four criterions listed. This number is 3 for Table 2. Thus, more or less there are some protein classes which are surely significant. Thus, it is more or less true that their definition of protein class is meaningful, i.e., not a random selection of proteins. On the other hand, there are some discrepancy among evaluations. However, most of the discrepancies occur in protein classes with a few proteins. Since clearly it is difficult to judge co ex-

A	B	C	D	ID	Annotation	N	d_{ij} D^N	d_{ij} P	$1-r_{ij}$ D^N	$1-r_{ij}$ P
○	○	○	○	1	ATP synthase	10	2.68e-02	1.30e-03	3.95e-01	1.40e-03
○	○	○	○	2	Ras superfamily	10	4.51e-02	4.90e-02	6.19e-01	1.64e-02
○	○	×	○	3	TCA cycle	22	4.82e-02	2.13e-02	7.95e-01	1.42e-02
○	○	×	○	4	adaptor	2	6.24e-03	4.97e-02	2.33e-02	3.19e-02
○	×	○	○	5	aldo-keto reductase superfamily	3	1.40e-02	4.06e-02	2.70e-01	8.36e-02
×	×	×	×	6	amino acid metabolism	19	5.90e-02	8.48e-01	9.96e-01	5.06e-01
○	○	○	○	7	antioxidant	16	4.90e-02	5.05e-02	7.37e-01	1.41e-02
○	○	×	×	8	apoptosis	2	2.59e-03	1.59e-02	6.58e-03	7.30e-03
○	○	○	×	9	beta-oxidation	8	3.66e-02	1.88e-02	6.08e-01	2.31e-02
×	○	×	×	10	calcium regulated protein	2	7.50e-03	6.66e-02	8.26e-03	9.50e-03
×	×	×	×	11	carbohydrate metabolism	18	5.06e-02	6.76e-02	8.63e-01	6.57e-02
×	×	×	×	12	cell-cell adhesion	2	1.69e-02	1.68e-01	4.66e-01	3.23e-01
×	×	○	×	13	chromatin-associated protein	3	4.27e-02	2.29e-01	6.61e-01	2.31e-01
○	○	○	○	14	cuticle	7	1.78e-02	2.30e-03	4.61e-01	1.29e-02
×	○	×	×	15	cytoskeleton	22	5.20e-02	9.01e-02	8.21e-01	2.38e-02
○	○	×	○	16	electron transport chain	15	2.67e-02	1.00e-04	3.60e-01	1.00e-04
○	○	○	○	17	energy storage	5	9.88e-03	1.10e-03	9.38e-02	2.20e-03
×	×	×	×	18	fatty acid synthesis	6	5.65e-02	3.91e-01	1.04e+00	5.22e-01
○	○	○	×	19	food	6	3.54e-02	3.46e-02	2.21e-01	3.30e-03
*	*	*	*	20	glycolipid metabolism	0				
×	×	○	×	21	helicase	3	3.45e-02	1.65e-01	5.90e-01	2.03e-01
×	×	○	×	22	hormone synthesis	4	5.13e-02	2.64e-01	6.81e-01	1.53e-01
×	×	×	×	23	immunity	2	9.58e-02	9.21e-01	1.86e+00	8.61e-01
×	×	×	○	24	kinases or phosphatases	2	9.27e-03	8.36e-02	3.34e-01	2.67e-01
×	×	×	×	25	membrane transporter	15	5.98e-02	8.80e-01	1.07e+00	9.70e-01
×	×	×	×	26	mevalonate pathway	3	2.89e-02	1.24e-01	3.57e-01	1.19e-01
×	×	×	×	27	nucleotide metabolism	2	6.31e-02	5.21e-01	4.53e-01	3.26e-01
×	×	○	×	28	pentose phosphate pathway	4	5.13e-02	2.66e-01	9.86e-01	3.85e-01
×	×	×	×	29	peptidase	15	6.00e-02	8.85e-01	1.01e+00	6.02e-01
*	*	*	×	30	peptidase inhibitor	1				
○	○	○	○	31	proteasome	24	2.47e-02	0.00e+00	3.12e-01	0.00e+00
×	×	×	○	32	protein folding	38	5.54e-02	2.76e-01	9.70e-01	3.14e-01
*	*	*	*	33	protein methylation	1				
×	×	○	×	34	protein receptor	4	6.37e-02	6.50e-01	9.87e-01	3.93e-01
○	○	×	○	35	ribonucleoprotein	4	2.28e-02	3.54e-02	2.51e-01	3.11e-02
○	○	×	○	36	ribosome	51	3.35e-02	0.00e+00	3.96e-01	0.00e+00
×	×	×	×	37	short-chain dehydrogenase family	14	5.18e-02	1.25e-01	9.41e-01	2.39e-01
×	×	×	×	38	small molecule carrier	7	6.16e-02	7.71e-01	1.03e+00	5.19e-01
×	×	○	×	39	small molecule receptor	4	4.67e-02	1.90e-01	7.73e-01	1.93e-01
×	×	○	×	40	tRNA synthetase	3	2.26e-02	8.64e-02	2.85e-01	9.57e-02
○	○	○	×	41	transcription	3	1.26e-02	3.41e-02	8.18e-02	1.74e-02
×	×	×	○	42	translation	15	5.18e-02	1.19e-01	8.33e-01	6.41e-02
○	○	○	○	43	ubiquitination	4	1.72e-02	2.01e-02	1.46e-01	1.33e-02
×	×	×	×	44	uncategorized	45	5.65e-02	5.31e-01	1.02e+00	9.48e-01
×	×	×	×	45	unknown function	23	5.92e-02	9.23e-01	1.05e+00	9.95e-01
×	×	×	×	46	vesicular transport	6	6.76e-02	9.84e-01	1.21e+00	9.88e-01

Table 1: Significance test for intra-protein class coexpressions for tissue samples. Significance due to A:D^N based upon nMDS, B:D^N based upon $1-r_{ij}$, C: Slope of expression trends (Table 3[6]), D: Enrichment in each node(Table 4[6]). ○: significant ($P < 0.05$ for A and B, as it is in Ref.[6] for C and D), ×: not significant, $*$ for missing. ID is numbers which are also used in the following figures. For details, see Materials and Methods.

A	B	C	D	ID	Annotation	N	d_{ij}		$1-r_{ij}$	
							D^{α}	P	D^{α}	P
*	*	*	*	1	ATP synthase	1				
×	×	×	×	2	Ras superfamily	3	9.99e-02	6.67e-01	9.32e-01	3.91e-01
×	×	○	×	3	TCA cycle	4	5.37e-02	9.45e-02	4.24e-01	7.70e-02
×	×	×	×	4	adaptor	2	7.35e-02	4.37e-01	6.26e-01	3.98e-01
×	×	×	×	5	aldo-keto reductase superfamily	3	4.20e-02	1.24e-01	3.67e-01	1.39e-01
×	×	○	×	6	amino acid metabolism	8	9.22e-02	8.55e-01	1.07e+00	8.65e-01
×	×	×	×	7	antioxidant	8	7.27e-02	1.45e-01	8.16e-01	1.72e-01
*	*	*	*	8	apoptosis	1				
*	*	*	*	9	beta-oxidation	1				
*	*	*	*	10	calcium regulated protein	1				
○	○	×	○	11	carbohydrate metabolism	15	6.94e-02	3.57e-02	7.28e-01	2.52e-02
*	*	*	*	12	cell-cell adhesion	1				
×	×	×	×	13	chromatin-associated protein	2	9.13e-02	5.42e-01	1.02e+00	5.46e-01
×	×	×	×	14	cuticle	2	1.18e-01	7.20e-01	1.64e+00	7.64e-01
×	○	×	×	15	cytoskeleton	7	7.22e-02	1.65e-01	5.71e-01	3.67e-02
*	*	*	*	16	electron transport chain	0				
○	○	○	○	17	energy storage	4	1.64e-02	3.50e-03	5.79e-02	4.20e-03
×	×	×	×	18	fatty acid synthesis	5	7.51e-02	2.60e-01	9.78e-01	4.80e-01
○	○	○	×	19	food	8	5.40e-02	1.89e-02	3.83e-01	6.70e-03
×	×	○	×	20	glycolipid metabolism	3	4.98e-02	1.51e-01	3.85e-01	1.43e-01
*	*	*	*	21	helicase	0				
*	*	*	*	22	hormone synthesis	0				
*	*	*	*	23	immunity	1				
*	*	*	*	24	kinases or phosphatases	0				
*	*	*	*	25	membrane transporter	0				
*	*	*	*	26	mevalonate pathway	0				
×	×	×	×	27	nucleotide metabolism	5	9.11e-02	6.52e-01	9.99e-01	5.16e-01
×	×	×	○	28	pentose phosphate pathway	5	5.55e-02	6.94e-02	5.24e-01	6.91e-02
○	○	×	○	29	peptidase	15	6.61e-02	2.09e-02	7.10e-01	2.26e-02
*	*	*	*	30	peptidase inhibitor	0				
○	○	○	○	31	proteasome	10	3.09e-02	2.00e-04	2.03e-01	5.00e-04
○	○	○	×	32	protein folding	20	5.45e-02	3.00e-04	5.70e-01	1.00e-03
*	*	*	*	33	protein methylation	0				
*	*	*	*	34	protein receptor	1				
×	×	×	×	35	ribonucleoprotein	2	5.12e-02	3.30e-01	4.49e-02	7.60e-02
○	○	○	○	36	ribosome	31	2.59e-02	0.00e+00	1.43e-01	0.00e+00
×	×	○	×	37	short-chain dehydrogenase family	4	7.84e-02	3.33e-01	9.76e-01	4.55e-01
×	×	×	×	38	small molecule carrier	3	8.03e-02	3.79e-01	9.61e-01	4.27e-01
×	×	○	×	39	small molecule receptor	4	4.31e-02	6.06e-02	3.17e-01	5.58e-02
*	*	*	*	40	tRNA synthetase	1				
×	×	×	×	41	transcription	2	1.56e-01	9.66e-01	1.15e+00	5.80e-01
○	○	○	×	42	translation	7	5.49e-02	3.22e-02	4.51e-01	1.87e-02
×	×	○	×	43	ubiquitination	4	8.29e-02	4.16e-01	9.02e-01	3.34e-01
×	×	×	○	44	uncategorized	21	7.77e-02	1.38e-01	8.81e-01	1.65e-01
×	×	×	×	45	unknown function	6	8.74e-02	5.79e-01	9.61e-01	4.56e-01
*	*	*	*	46	vesicular transport	1				

Table 2: The same analysis as Table 1 for hemolymph

pression within only a few proteins, thus we do not discuss about these cases here. What we would like to emphasize here is that our method, in principal, can judge if classes with only two proteins (e.g., adaptor (ID 2), aptosis (ID 8) et cetera in Table 1) co express or not, which is hard by the method in Ref. [6].

Notable discrepancies for protein classes with more proteins include Ribosome (ID 36) for tissue samples. Our methods and slope criterion (Table 3[6]) evaluated this to be co expressed significantly, but cluster method (Table 4[6]) did not. The reason can be seen in Figs. **??**. For tissue samples, proteins in the Ribosome protein class are scattered. This is simply because ribosome protein class include both of proteins with positive/negative slopes against time. In such cases, it is very hard to be clustered together since positive and negative slop cluster are major two clusters in tissue experiment. Due to that, cluster analysis[6] cannot cluster them together, thus fail to detect. Clearly, our method can detect this missing co expression of proteins in Ribosome protein class.

It is apparent that our method can figure out more hidden relationship between protein classes than conventional cluster method[6]. In the cluster method, proteins are first clustered then tested if being significant. Thus, if the clustering is not appreciated, no significance can be detected. The method without assuming pre-existence of cluster, like ours, can detect significant co expression of proteins within ribosome class. In addition to this, we can give universal criterion (e.g., P-value) for the definition of co expression, which laked in cluster investigation. In cluster analysis, it is unclear what the co expressed set of proteins is. Because of these reasons, our method can depict biologically meaningful coexpression clusters of protein classes.

The reason why there are many nodes which are connected with only either or both of proteasome and ribosome classes in Figs. 1 is because if the class is too small we cannot exclude the possibility that they are coexpressed (see Methods and Materials). If we group large and well-localized class with small classes, their overall distribution can differ from uniform one, thus are regarded as coexpressed. In order to deny this possibility, we need more proteins annotated in the smaller class.

We have also confirmed that configuration converged is almost free from the initial configurations from which iteration starts (not shown here). Thus, we believe that our results obtained by nMDS is more suitable than that by conventional clustering methodology.

4 Conclusion

In this paper, we have applied nMDS to protein expression profiles and have found that proteins in many protein classes can be regarded as being coexpressed. By applying significance test to pairs of protein classes, we have obtained coepxression network of proteins, which turns out to be very reasonable and informative.

5 Materials and Methods

We have employed five days gene expression profiles of proteins[6]. These are included both or either of tissue samples (475 proteins) and hemolymph samples (222 proteins), since those with at least four days profile are used. These are also annotated by more than one of the 47 protein classes.

In order to see visual inspection between these gene expression profiles, we have employed nMDS[7] to embed gene expression profile of each protein into two dimensional space. To do this, we have used negative signed correlation coefficients between gene expression profiles as dissimilarity. In nMDS embeddings, rank order of distances between points which express each gene expression profile of protein is tried to match with rank order of dissimilarities. The significance of embeddings of each point (protein) can be checked by test that rank order of both distance and dissimilarity attached to each

point match significantly better than random configurations. In this case, very few number of points have larger P-values (> 0.05),thus we can regard that these embeddings are good enough.

Then we have defined significance of intra class proteins co expression as follows. Say, there are N proteins within a protein class. First we compute test variable as

$$D^N \equiv \sum_{ij} d_{ij},$$

where d_{ij} is the distance between points within the protein class. Then we have picked up the same number of proteins randomly 10,000 times and computed the distribution test variables under the null hypothesis. Then P-value is the possibility that the distribution takes smaller values than the above computed value.In order to see if two protein class is significantly coexpressed, we have done almost same.

If two protein classes include N and M proteins respectively, first, we compute

$$D^{N+M} \equiv \sum_{ij} d_{ij},$$

where summation is taken over any pairs taken from $N+M$ proteins which belong to either of two classes. Then we have computed same variables for randomly selected $N + M$ proteins 10,000 times. P-value is computed as the same as above. However, in this criterion, closely related two classes can be regarded as being coepxressed. In order to avoid this possibility, we compute

$$\frac{D^{N+M}}{\sqrt{D^N D^M}}$$

and compute P value that this variable under null hypothesis is higher than this. If this P value is smaller than 0.5, we can deny the possibility that these two classes are truly co expressed. Thus, we excluded such pairs of classes from pairs of co expressed protein classes. It is expected that expectation of D^{N+M}, D^M, and D^N take the same value if all $N + M$ proteins belong to the same cluster and distribution within cluster is the same. However, distribution in each differs from each other, the expectation of this ratio can differ from 1.

If P value by D^{N+M} is less than P_0 and that by $\frac{D^{N+M}}{\sqrt{D^N D^M}}$ is not less than 0.5, we regard this pair of protein classes as being co expressed. P_0 for tissue and hemolymph is 0.05. This is decided as follows. As P_0 increases, significance of coepxression decreases. In order to check its significance, we have repeated the same procedure using the distance computed from dissimilarity,

$$1 - r_{ij}$$

where r_{ij} is the correlation coefficients between gene expressions. Then we check the coincidence between significant pairs proteins by two distances d_{ij} and $1 - rij$ by changing P_0. Then we find that P_0 exceeds 1×10^{-2}, coincidence has saturated. Then, we employ $P_0 = 1 \times 10^{-2}$ as lower limit of reliable threshold values. Since 10 to 20 % of edges should be drawn for informative graphs, we have employed these values.

A protein coexpression network graph, where node represent each protein class and edge represents co expression pair of protein class, is drawn by plot command for graph object in the R package igraph assuming Fruchterman and Reingold method.

6 Acknowledgements

This work has been supported by the Grant-in-Aid for Creative Scientific Research No.19500254 of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2007 to 2008. We are grateful for their support.

References

[1] Consortium THGS, Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* 2006, **443** pp.931-949.

[2] Chan QW, Howes CG, Foster LJ, Quantitative comparison of caste differences in honeybee hemolymph, *Mol. Cell Proteomics* 2006, **5**, pp. 2252-2262.

[3] Scharlaken B, de Graaf DC, Goossens K, Peelman LJ, Jacobs FJ, Differential gene expression in the honeybee head after a bacterial challenge, *Dev. Comp. Immunol.* 2008, **32** pp.883-889.

[4] Wolschin F, Amadam GV, Comparative proteomics reveal characteristics of life-history transition in a social insect. *Proteome Sci.* 2007, **5** p.10.

[5] Wolschin F, Amadam GV, Plasticity and robustness of protein patterns during rversible development in the honey bee (Apis mellifera). *Anal. Bioanal. Chem.* 2007, **389** pp.1095-1100.

[6] Chan QW, Foster LJ, Changes on protein expression during honey bee larval development. *Genome Biology* 2008, **9** R156.

[7] Taguchi Y-h, Oono Y, Relational patterns of gene expression via nonmetric multidimensional scaling analysis. *Bioinformatics*, 2005, **21** pp.730-40.