

Estimation of Metabolic Effects of Cadmium Exposure during Pregnancy by Tensor Decomposition

YUKI AMAKURA^{1,a)} Y-H TAGUCHI^{1,b)}

Abstract:

studies have shown that exposure to cadmium during maternal pregnancy disrupts insulin metabolism in the unborn child. The purpose of this study is to estimate the genes actually affecting by the data obtained in this report using tensor decomposition and to speculate on the possible mechanisms of insulin metabolism disruption by these genes. By identifying genes with high expression levels, we were able to speculate on a pathway in which the effects of endoplasmic reticulum stress induced by cadmium might somehow promote increased expression of TNF- α , and the effects of intracellular oxidative stress via TNF- α -induced ceramide accumulation might lead to phosphorylation of IRS1, which plays an important role in insulin signaling. We hypothesized a pathway in which insulin resistance is acquired by phosphorylation of a protein, IRS1.

1. Introduction

Cadmium is a non-essential trace metal element in cells¹⁾. However, when excessive amounts of cadmium are taken into the body due to cadmium inoculation by food or cadmium exposure, cadmium accumulates in the cells. Because cadmium is an electrophilic metal, it has a negative effect by acting on protein molecules in the cell¹⁾. One of the most prominent effects is inhibition of protein folding. In order for a protein to fully perform its function, it is important that its three-dimensional structure is maintained properly, and unfolding proteins accumulate in the endoplasmic reticulum, which is responsible for folding proteins. The accumulation of this unfolding protein is known to cause endoplasmic reticulum stress, and excessive stress induces endoplasmic reticulum-associated degradation (ERAD). When ERAD is induced, unfolding proteins are degraded by the ubiquitin proteasome system. In other words, accumulated unfolding proteins undergo ubiquitination, making them targets for degradation by the proteasome, and as degradation proceeds, endoplasmic reticulum stress is relieved. However, when the stress accumulates beyond the stress avoidance capacity of ERAD, cells induce apoptosis. ERAD also promotes ATP-dependent responses. In addition, insulin signaling is initiated by the binding of insulin to the insulin receptor (IR) on the cell membrane, which ultimately moves GLUT4 to the cell membrane, thereby bringing glucose into the cell. The glucose taken up by the body is used by the glycolytic system in the cell to produce ATP. When insulin signaling is inhibited, symptoms such as elevated blood sugar levels are seen, which is also the

cause of type 2 diabetes.

We already know about the effects of cadmium on cells, but we do not know the specific mechanism of how this effect on cells affects insulin signaling.

In the analysis of the present data, we will identify the genes involved in this phenomenon and infer the specific mechanism from these genes.

2. Dataset used in the analysis

2.1 Preparation of the data set

"GSE150679"^{*1} was downloaded from Gene Expression Omnibus (GEO), a database provided by the National Center for Biotechnology Information (NCBI), USA.

This dataset contains the presence or absence of maternal exposure to cadmium, the sex of the offspring mice, the biological replicates for each of these conditions, and the gene expression levels of the offspring mice under these conditions. This data set also consists of three Exel files, each consisting of data from day 1 (PND1), day 21 (PND21), and day 42 (PND42) after birth.

2.2 Data Processing

The data were formatted into tensors for use in tensor analysis. First, sex differences were omitted from this analysis because there were not enough data from conditionally differentiated littermates to formulate a tensor (This is because only the PND42 file has conditional distinctions for gender, and these distinctions were not enough to form a tensor with the PND1 and PND21 data.)

Next, for the biological replicates, since the number of individuals under the same conditions differed among the three files, three data sets of individuals under the same conditions

¹ Chuo University Faculty of Science and Engineering

^{a)} a17.rtps@g.chuo-u.ac.jp

^{b)} tag@granular.com

^{*1} <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150679>

were selected for analysis. Specifically, the columns of the Exel file were rearranged in the following order. The order is clearly indicated by the sample ID.

PND1 : GSM4556351 GSM4556352 GSM4556353
GSM4556347 GSM4556348 GSM4556349
PND21: GSM4556358 GSM4556359 GSM4556360
GSM4556355 GSM4556356 GSM4556357
PND42: GSM4556365 GSM4556366 GSM4556367
GSM4556361 GSM4556362 GSM4556363

The number of genes whose gene expression levels were commonly obtained in the three files was 12795, so the analysis was focused on these genes (Only PND42 contained data on the expression levels of more genes than the other two.).

2.3 Tensor used in the analysis

2.3.1 First dimension of the tensor: Gene

As described above, 12795 genes were analyzed.

2.3.2 The second dimension of tensor: Biological Replicates

As described above, we selected three data sets of individuals under the same conditions.

2.3.3 Third dimension of tensor: days old (time lapse)

The order of birth was 1 day, 21 days, and 42 days, which was the third dimension of the tensor.

2.3.4 Fourth dimension of tensor: presence of maternal cadmium exposure

We analyzed whether there is a difference in gene expression in the offspring depending on maternal exposure to cadmium.

2.3.5 Rank of the tensor

We have created a fourth-order tensor x_{ijkl} with these variables, and the parameters representing the rank of the matrix in each dimension are $(N, M, K, L) = (12795, 3, 3, 2)$.

3. How to analysis

3.1 Unsupervised learning using tensor decomposition

In this analysis, we use the Tucker decomposition among the tensor decompositions, and Higher Order Singular Value Decomposition (HOSVD²⁾ is used as the algorithm.

As an example, there is a third-order tensor x_{ijk} with ranks N , M , and K in each dimension (The tensor treated in this analysis is a fourth-order tensor, but it is easy to extend the Tucker decomposition from a third-order tensor to a fourth-order tensor.). The Tucker decomposition can then be performed on this tensor, using the orthogonal matrix U , and decomposed as follows

$$x_{ijk} = \sum_{l_1}^N \sum_{l_2}^M \sum_{l_3}^K G(l_1, l_2, l_3) u_{l_1 i} u_{l_2 j} u_{l_3 k}$$

The Tucker decomposition decomposes the original data tensor into the core tensor $G(l_1, l_2, l_3)$ and three singular value matrices, $u_{l_1 i}, u_{l_2 j}, u_{l_3 k}$. These are the matrices that represent the dependencies on variables i , j , and k , respectively.

Suppose we want to find a variable k with a large value

in matrix $u_{l_3 k}$ that is of interest for the purpose of analysis. The first step is to select the parameters l_1' and l_2' that follow some dependency of interest in $u_{l_1 i}$ and $u_{l_2 j}$, respectively (For example, if i is a variable that represents time variation, it would be preferable that the vector $u_{l_1 i}$ for the chosen l_1' be a straightforward vector such as monotonic increase or decrease.).

Next, substitute the selected l_1' and l_2' into the core tensor $G(l_1, l_2, l_3)$. Since each element of the core tensor represents a weight for the product of each vector specified by the parameter (l_1, l_2, l_3) , the parameter l_3 that has a large absolute value in the resulting vector $G(l_1', l_2', l_3')$, which has a large absolute value in the vector $G(l_1', l_2', l_3')$ thus obtained, has a large contribution to the original tensor. Based on this idea, we choose l_3' .

Next, we obtain the vector $u_{l_3' k}$ by substituting l_3' into $u_{l_3 k}$. This vector follows the dependency of the chosen parameters l_1', l_2' and has a large contribution to the original tensor.

3.2 Variable selection by χ -square test

Finally, χ -square test is performed using the normalized version of this vector to find the k with a particularly large value among the elements of the obtained vector $u_{l_3' k}$, and a P-value is assigned.

$$p_k = p_{\chi^2} \left[> \sum_{l_3'} \left(\frac{u_{l_3' k}}{\sigma_{l_3'}} \right)^2 \right]$$

The above objective is achieved by applying further multiple comparison correction to it to obtain k with a p-value less than 0.05.

4. Result

4.1 Identification of genes with large expression levels

When we decompose the fourth-order tensor used in this analysis in R by HOSVD, we obtain a list of Z that contains the elements of the core tensor and a list of U that contains the four singular value matrices.

The tensor used is as described above, where subscript i is the gene, j is the biological replicate, k is the time course, and l is the variable for the presence or absence of maternal exposure to cadmium. Since the purpose of this analysis is to identify which gene i has a large expression level due to the effect of cadmium exposure, if we can find a parameter l_1 in the matrix $u_{l_1 i}$ that is dependent on cadmium exposure, invariant to biological replicates, and somehow time-dependent. If we can find a parameter l_1 in the matrix $u_{l_1 i}$ that is equivalent to the dependence on cadmium exposure, invariance to biological replicates, and some time dependence, we can assign that l_1 to $u_{l_1 i}$ to obtain the gene i of interest. To do this, we start by selecting l_2 , l_3 , and l_4 .

First, we took $u_{l_2 j}$ from object U . We chose $l_1 = 1$ because it is a biological replicate, that is, the smaller the individual difference under the same conditions, the better (Fig. 1).

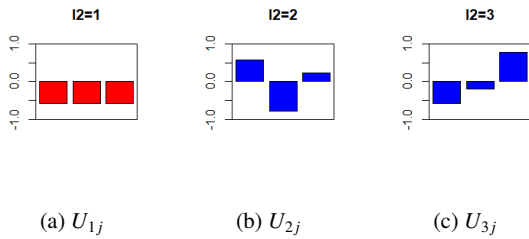


Fig. 1: U_{l_2j}

Next, we took out the u_{i_3k} . We chose $l_3 = 2$, which follows a monotonic decrease, because extracting a time dependence that follows a monotonic increase or decrease leads to a better interpretation of whether the expression level is increasing or decreasing as days pass from birth (Fig. 2).

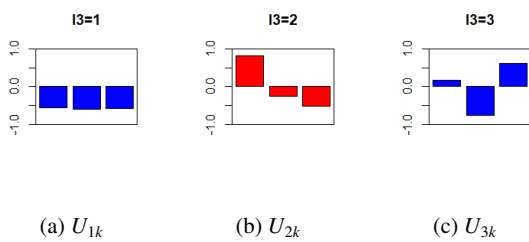


Fig. 2: U_{l_3k}

Next, we took out u_{i_4l} . Since the purpose of this study is to identify genes that are differentially expressed in the presence or absence of cadmium exposure, $l_4 = 2$ was selected (Fig. 3).

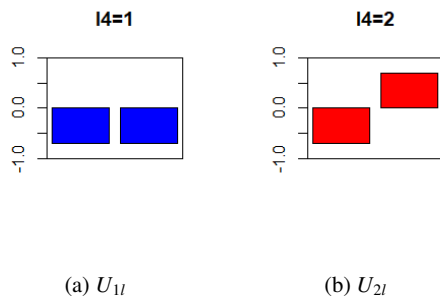


Fig. 3: U_{l_4l}

Now since we have selected $l_2 = 1$, $l_3 = 2$, and $l_4 = 2$, we are ready to identify genes that have small individual differences under the same conditions, that differ with respect to maternal exposure to cadmium, and that are expressed in decreasing amounts over time.

Next, by taking the list of objects Z and specifying $G(l_1, 1, 2, 2)$ and parameters in the core tensor G , a vector of length 12795 is obtained. The element specified by l_1 , which has a large absolute value on this vector, follows the dependency specified by the parameters $l_2 = 1$, $l_3 = 2$, $l_4 = 2$, and has a large contribution to the original tensor. Due to the dimensionality of the original tensor, the 19th and subsequent elements on the vector $G(l_1, 1, 2, 2)$ were very small, so the values of the elements from 1 to 18 are shown in Fig. 4. We chose $l_1 = 3, 11, 14$ with

large absolute values as the values of l_1 to be assigned to u_{l_1i} . These were substituted into u_{l_1i} to obtain the three vectors.

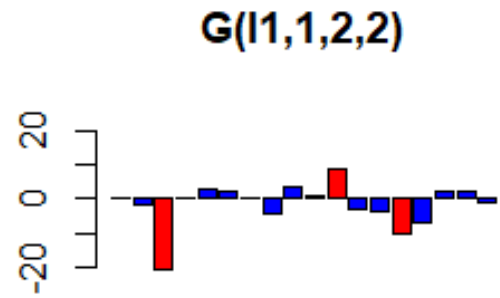


Fig. 4: $G(l_1, 1, 2, 2)$

To identify genes with particularly large expression levels, we assume that u_{3i} , u_{11i} , and u_{14i} independently follow a normal distribution, and assign P-values by the χ -square test. Using these three vectors, execute the `p.adjust` command.

In addition, the P-value given above for each gene i is corrected for multiple comparisons using the `p.adjust` function, and then the i whose P-value is less than 0.05 is selected. The final result is that the gene i chosen here follows the dependency of interest and is a gene with a very large expression level. In this analysis, 204 genes were selected from 12795 genes.

4.2 Checking the ontology with gProfiler

“gProfiler”^{*2} is a web tool developed by the Estonian research group Bioinformatics, Algorithmics and Data Mining Group (BIIT). Enrichment analysis was performed on the 204 genes obtained in the analysis, and the following ontologies were hit as shown in the table below.

Database	Term Name	Term ID	P-Value
KEGG	Metabolic pathway	KEGG01100	1.661E-6
KEGG	Activation of platelets	KEGG04611	2.294E-2
REAC	Lipid metabolism	R-HSA-556833	1.107E-2
GO BP	Ionic transport	GO0006811	2.741E-8
GO BP	Cell adhesion	GO0007155	4.702E-6
GO BP	Intracellular signaling	GO0035556	4.207E-6
GO BP	Cellular protein modification process	GO0036211	5.519E-9
GO BP	Regulation of gene expression	GO0010468	1.241E-8
GO BP	RNA metabolism	GO0016070	9.895E-11
GO CC	Mitochondria	GO0005739	1.990E-4
GO CC	Endoplasmic reticulum	GO0005783	1.990E-3

*2 <https://biit.cs.ut.ee/gprofiler/>

5. Discussion

5.1 Functions of the identified gene

The purpose of this study was to estimate the pathway by which insulin signaling is inhibited. To do so, we first checked whether the genes involved in ERAD are highly expressed.

As mentioned above, during ERAD development, protein degradation is promoted by the ubiquitin-proteasome system. Among the genes obtained were UBE2E3 with E2 ubiquitin ligase activity³⁾, AMFR with E3 ubiquitin ligase activity⁴⁾, and others with ubiquitin ligase activity such as E4F1⁵⁾ and TRIM32⁶⁾. There was also COMMD9, a regulator of these activities⁹⁾. Genes involved in the proteasome included PSMC2, which encodes a subunit of the 26s proteasome⁸⁾, and PSME1, which is part of the proteasome activator PA28⁹⁾. This high expression of multiple genes involved in the ubiquitin-proteasome system may be responsible for the development of ERAD.

The genes involved in NF- κ B, which plays a central role in the immune response, were TBK1, which indirectly interacts with NF- κ B¹⁰⁾, CARD11, which is an activator¹¹⁾, and COMMD9, which is a negative regulator¹²⁾, suggesting that NF- κ B may be highly expressed. It is activated by the cytokine TNF- α , and since high expression of NFAM1, which encodes a receptor that activates the promoter of TNF- α ¹³⁾, was observed, it is possible that TNF- α is also highly expressed. Studies have shown that there is a positive correlation between TNF- α expression and the development of ERAD¹⁴⁾, and the results of this study also suggest this positive correlation. Given that ERAD proceeds in an ATP-dependent manner, we checked to see whether any genes that upregulate ATP production were highly expressed, and found ATP5G1, which encodes a subunit of the enzyme that catalyzes ATP production.

The genes involved in the electron-transfer system were NDUFAF7, which is involved in stabilizing the assembly of electron-transfer complex enzymes¹⁵⁾, and ETFBKMT, which may negatively regulate electron-transfer system function¹⁶⁾. DLAT, which plays a role in linking the electron-transfer system to the glycolytic system¹⁷⁾, was also highly expressed. Metabolites of glycolysis and fatty acid metabolism are sent to the electron transport system to promote ATP production in mitochondria. The genes involved in fatty acid metabolism were GPIHBP1, which promotes fatty acid degradation¹⁸⁾, and PPARA, which encodes a growth factor-activated receptor for peroxisomes¹⁹⁾, the site of fatty acid metabolism. The high expression of genes related to glycolysis and fatty acid metabolism may well be due to the activation of the electron-transfer system, considering that these metabolites are passed on to the electron-transfer system. Activation of the electron-transfer system directly leads to activation of ATP production.

In addition, SPTLC2²⁰⁾ and SAMD8²¹⁾, which are involved in the biosynthesis of sphingolipids, were highly expressed. This suggests that sphingolipids may be over-synthesized in the cell.

5.2 Inference of the mechanism of insulin metabolism inhibition from the obtained genes

The mother's exposure to cadmium causes the fetus to be exposed to cadmium as well²²⁾. Excessive cellular uptake of cadmium causes inhibition of protein folding, and this leads to the development of ERAD. The high expression of genes related to the ubiquitin-proteasome system and genes related to glycolytic, fatty acid metabolism, and electron transfer systems in this analysis can be considered to promote ATP synthesis, which may support the development of ERAD. The positive correlation between the expression levels of ERAD and TNF- α , and the high expression of genes related to TNF- α and NF- κ B, suggest that TNF- α may be over-secreted in the cells. TNF- α is known to activate ASMase (oxidized sphingomyelinase), and ASMase promotes the production of sphingolipids. In this analysis, high expression of genes involved in the production of sphingolipids was also observed, suggesting that accumulation of sphingolipids may be occurring. Accumulation of ceramide, a type of sphingolipid, is known to lead to overproduction of ROS (Reactive Oxygen Species). In addition, ROS is produced in the reaction process of the electron transfer system. The effect of oxidative stress by ROS activates ASK1²³⁾. Among the genes that were found to be highly expressed in this analysis was PISD. It also plays a role in the activation of ASK1. ASK1 activates JNK, and JNK inactivates it by phosphorylating 307 serine residues of IRS1, which are essential in insulin signaling. Thus, a pathway for the acquisition of insulin resistance due to the effects of oxidative stress via TNF- α -induced ceramide accumulation can be inferred from the results of this analysis. However, we could not speculate on the mechanism of the positive correlation between ERAD and high expression of TNF- α from this analysis.

In addition to the genes mentioned above, there were also hits for ontologies such as "regulation of gene expression" and "RNA metabolism" as shown in the table. This suggests that some kind of change in gene expression is occurring, but the specific mechanism could not be inferred in this study.

We also found a number of genes related to apoptosis, which is a last resort for stress avoidance in ERAD development, but it is unclear how apoptosis is actually related to the pathway of insulin resistance acquisition, which we speculated in this study.

Future work will focus on understanding how these factors are involved, and in particular, what causes and influences changes in gene expression.

References

- [1] Ken-ichi Ohba. Transport and Toxicity of Cadmium, *Nichieishi*, Vol. 73, pp. 269-274 (2018).
- [2] Y-h Taguchi. Unsupervised Feature Extraction Applied to Bioinformatics : A PCA Based and TD Based Approach. Springer International Publishing AG. (2019) ,329p.
- [3] Ito K., Kato S., Matsuda Y., Kimura M., Okano Y. cDNA cloning, characterization, and chromosome mapping of UBE2E3 (alias UbcH9), encoding an N-terminally extended human ubiquitin-conjugating enzyme, *Cytogenet Cell Genet*, Vol. 84, pp. 99-104 (1999).
- [4] Marilyn Registre., Jacky G. Goetz., Pascal St. Pierre., Hao Pang., Monique Lagace., Michel Bouvier., Phuong U.Le., Ivan R. Nabi., The gene product of the gp78/AMFR ubiquitin E3 ligase cDNA is selectively recognized by the 3F3A antibody within a subdomain of the endoplasmic reticulum, *Biochemical and Biophysical Research*

- Communications, Vol. 320, No. 4, pp. 1316-1322 (2004).
- [5] Laurent Le Cam., Laëtitia K. Linares., Conception Paul., Eric Julien., Matthieu Lacroix., Elodie Hatchi., Robinson Triboulet., Guillaume Bossis., Ayelet Shmueli., Manuel S. Rodriguez., Olivier Coux., Claude Sardet., E4F1 Is an Atypical Ubiquitin Ligase that Modulates p53 Effector Functions Independently of Degradation, *Cell*, Vol. 127, No.4, pp.775-788 (2006).
- [6] Annie P. Chiang., John S. Beck., Hsan-Jan Yen., Marwan K. Tayeh., Todd E. Scheetz., Ruth E. Swiderski., Darryl Y. Nishimura., Terry A. Braun., Kwang-Youn A. Kim., Jian Huang., Khalil Elbedou.r, Rivka Carmi., Diane C. Slusarski., Thomas L. Casavant., Edwin M. Stone., and Val C. Sheffield., Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11), *PNAS*, Vol. 103, No.16, pp. 6287-6292 (2006).
- [7] Weihua Zhan., Weanjuan Wang., Tianyu Han., Caifeng Xie., Tingting Zhang., Mingxi Gan., Gian-Bin Wang., COMMD9 promotes TFDP1/E2F1 transcriptional activity via interaction with TFDP1 in non-small cell lung cancer, *Cellular Signalling*, Vol. 30, pp. 59-66 (2017).
- [8] Nobuyuki Tanahashi., Mikio Suzuki., Tsutomu Fujiwara., Ei-ichi Takahashi., Naoki Shimbara., Chin Ha Chung., Keiji Tanaka., Chromosomal Localization and Immunological Analysis of a Family of Human 26S Proteasomal ATPases, *Biochemical and Biophysical Research Communications*, Vol. 243, No. 1, pp. 229-232 (1998).
- [9] C.P.Ma., C.A.Slaughter., G.N.DeMartino., Identification, purification, and characterization of a protein activator (PA28) of the 20 S proteasome (macropain), *Journal of Biological Chemistry*, Vol. 267, No. 15, pp. 10515-10523 (1992).
- [10] J L Pomerantz., D Baltimore., NF-kappaB activation by a signaling complex containing TRAF2, TANK and TBK1, a novel IKK-related kinase, *EMBO Journal*, Vol. 18, No. 23, pp. 6694-6704 (1999).
- [11] John Bertin., Lin Wang., Yin Guo., Michael D. Jacobson., Jean-Luc., Poyet Srinivasa M. Srinivasula., Sarah Merriam., Peter S. Di Stefano., Emad S. Alnemri., CARD11 and CARD14 Are Novel Caspase Recruitment Domain (CARD)/Membrane-associated Guanylate Kinase (MAGUK) Family Members that Interact with BCL10 and Activate NF- κ B, *Journal of Biological Chemistry*, Vol. 276, No.15, pp. 11877-11882 (2001).
- [12] G. N. Maine., E. Burstein., COMMD proteins: COMMMing to the scene, *Cellular and Molecular Life Sciences*, Vol. 64, pp. 1997-2005 (2007).
- [13] Martha L Slattery., Abbie Lundgreen., Kristina L Bondurant., Roger K Wolff., Tumor necrosis factor-related genes and colon and rectal cancer, *International Journal of Molecular Epidemiology and Genetics*, Vol. 2, No. 4, pp 328-338 (2011).
- [14] Xiaoqing Yang., Qingqing Liu., Guangyu Long., Yabin Hu., Endoplasmic Reticulum Associated Degradation(ERAD) Deficiency Promotes Mitochondrial Dysfunction and Transcriptional Rewiring in Human Hepatic Cells, *Journal of Biological Chemistry*, Vol. 295, pp. 16743-16753 (2020).
- [15] Olga Zurita Rendón., Lissiene Silva Neiva., Florin Sasarman., Eric A. Shoubridge., The arginine methyltransferase NDUF7 is essential for complex I assembly and early vertebrate embryogenesis, *Human Molecular Genetics*, Vol. 23, No. 19, pp. 5159-5170 (2014).
- [16] Jędrzej Malecki., Magnus E. Jakobsson., Angela Y. Y. Ho., Anders Moen., Arild C. Rustan., Pål Ø. Falnes., Uncovering human METTL12 as a mitochondrial methyltransferase that modulates citrate synthase activity through metabolite-sensitive lysine methylation, *Journal of Biological Chemistry*, Vol. 292, No.43, pp. 17950-17962 (2017).
- [17] Berg JM., Tymoczko JL., Stryer L. New York., W H Freeman. *Biochemistry* (5th edition), Elsevier India. (2020), 794p. (2002)
- [18] Young Stephen G., Davies Brandon SJ., Fong Loren G., Gin Peter., Weinstein Michael M., Bensadoun André., Beigneux Anne P., GPIHBP1: an endothelial cell molecule important for the lipolytic processing of chylomicrons, *Current Opinion in Lipidology*, Vol. 18, No.4, pp. 389-396 (2007).
- [19] Sander Kersten., Integrated physiology and systems biology of PPAR α , *Molecular Metabolism*, Vol. 3, No. 4, pp. 354-371 (2014).
- [20] Akio Kihara., Various Metabolisms., Physiological Functions and Related Diseases of Fatty Acids, *Biochemistry*. Vol. 82, No. 7, pp. 591-605 (2010).
- [21] Fikadu G. Tafesse., Ana M. Vacaru., Elleke F. Bosma., Martin Hermansson., Amrita Jain., Angelika Hilderink., Pentti Somerharju., Joost C. M. Holthuis., Sphingomyelin synthase-related protein SMSR is a suppressor of ceramide-induced mitochondrial apoptosis, *Journal of Cell Science*, Vol. 127, No. 2, pp. 445-454 (2014).
- [22] Atsuhiko Nakano., Masaji Ono and Yukiko Ohta., Ichiro Wakisaka., Placental Transfer of Cadmium in Pregnant Women, *Japanese Journal of Hygiene*, Vol. 44, No. 2, pp. 579-586 (1989).
- [23] Tsuruzoe Kaku , Nishikawa Takesi , Araki Eichi., Oxidative Stress and Insulin Resistance, *Diabetes*, Vol. 49, No.11, pp. 845-848 (2006).