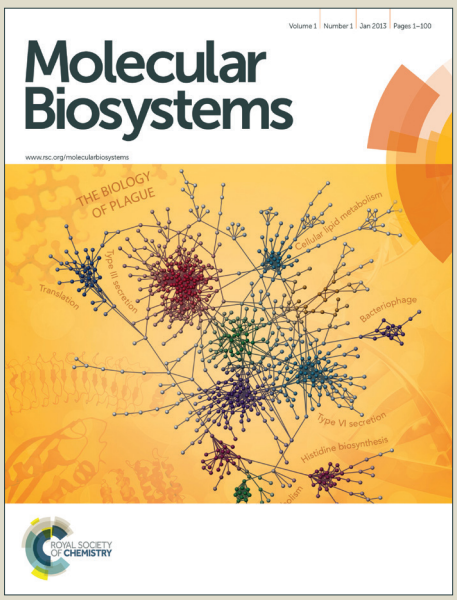
Molecular BioSystems

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model

Junli DU[1, 2], Zhifa YUAN [1,*], Ziwei MA[1], John SONG[3], Xiaoli XIE[1], Yulin CHEN [2,*]

([1]College of sciences, Northwest A&F University, Yangling, 712100, P. R. China)

([2]College of Animal Science and Technology, Northwest A&F University, Yangling, 712100, P. R. China)

([3]Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA)

*Corresponding author; E-mail: liml75@126.com; chenyulin@nwafu.edu.cn

Molecular BioSystems Accepted Manuscript

Abstract: The dynamic impact approach (DIA) represents an alternative to overrepresentation analysis (ORA) for functional analysis of time-course experiments or those involving multiple treatments. The DIA can be used to estimate the biological impact of the differentially expressed genes (DEG) associated with particular biological functions, for example, as represented by Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations. But the DIA does not take into account the correlated dependence structure of the KEGG pathway hierarchy. We have developed herein a Path analysis model (KEGG-PATH) to subdivide the total effect of each KEGG pathway into direct effect and indirect effect by taking into account not only each KEGG pathway itself, but also the correlation with its related pathway. In addition, this work also attempts to preliminarily estimate the impact direction of each KEGG pathway by the gradient analysis method from principal component analysis (PCA). As a result, the advantage of KEGG-PATH model is demonstrated through the functional analysis of bovine mammary transcriptome during lactation.

Keywords: Dynamic impact approach; Path analysis model; Gradient analysis; Principal component analysis; Bovine mammary

Molecular BioSystems Accepted Manuscript

1. Introduction

In order to identify the most relevant pathways in a given experiment, the functional analysis of genomics data--the substantial genes, especially differentially expressed genes (DEG), produced by high-throughput genome-wide experiments--is often conducted by the enrichment analysis, also called overrepresented approach or ORA [1-4]. During the last few decades, there are approximately over 68 bioinformatics enrichment tools using this ORA approach [3]. In spite of its wide adoption, this approach has a number of limitations related to the type, quality, and the structure of the annotations available [1]. Especially for the time-course experiments or experiments involving multiple treatments, the ORA do not allow comparing the results from the different experimental conditions [4].

Recently, a novel dynamic impact approach (DIA) was first proposed for functional analysis of time-course experiments or those involving multiple treatments and had been successfully validated using microarray data from a large time-course experiment of bovine mammary tissue during an entire lactation cycle [4]. The DIA approach aims to calculate the impact value and the impact direction of the biological terms (pathways/functions). Three factors are considered in the calculation: the percentage of the DEG vs. the total genes assigned to the term, magnitude of the change of DEG and average significance of DEG. It is important that the total genes assigned to the pathway/function are considered, which can increase biological relevance of the results. However, the DIA method fails to exploit how various

Molecular BioSystems Accepted Manuscript

pathways interact with each other. In fact, Kyoto Encyclopedia of Genes and Genomes (KEGG) is a complex network structure including the KEGG pathway categories, subcategories and the secondary pathways, which mainly describe metabolic pathway and gene signaling networks [5]. These pathways in the cell are highly interconnected. Still further, there is strong regulation mechanisms existed in these pathways. For example, in metabolic pathways, quite often the product of one pathway becomes the substrate for another [6]. Therefore, it is essential to discover the regulation mechanisms and principles that underlie cell function among the KEGG pathways.

Here we proposed a Path analysis model to deeply explore the regulating mechanisms among the KEGG pathways based on the DIA impact value. This method tends to subdivide the total effect of a specified KEGG pathway into direct effect and indirect effect on it by Path analysis method. And the total effect can be used to measure the importance of its biological impact. In addition, we also attempt to preliminarily estimate the impact direction of each KEGG pathway by the gradient method from principal component analysis (PCA). The utility of these methods are demonstrated using the DIA impact value dataset from the functional analysis of the bovine mammary transcriptome during the lactation cycle [6].

2. Method

2.1 The Path analysis model of KEGG pathway

Molecular BioSystems Accepted Manuscript

The proposed method is called KEGG-PATH: KEGG-based pathway analysis using a Path analysis model. In this model, each KEGG pathway has three relevance parameters measuring its direct effect, indirect effect and total effect. The novelty of the model is that the total effect of a specified KEGG pathway is subdivided into direct effect and indirect effect. This subdivision can fully demonstrate the regulating mechanisms of the KEGG pathways. Meanwhile, the total effect can be used to measure the importance of each KEGG pathway.

To introduce KEGG-PATH, we define the following notations.

We assumed that $X=(X_1,X_2,\cdots,X_m)^T, X_i=(X_{i1},X_{i2},\cdots,X_{ip})^T,(i=1,2,\cdots,m)$, $X_{ij}=(X_{ij1},X_{ij2},\cdots,X_{ijk})^T,(i=1,2,\cdots,m;j=1,2,\cdots,p)$ are the sets of KEGG pathway categories, sub-categories and the secondary pathways, respectively. Then, the network data structure of the KEGG pathway can be illustrated as Figure 1.

The path analysis will be made between KEGG pathway categories and sub-categories, and between the subcategories and its secondary pathways.

The Path analysis between KEGG pathway categories and subcategories was used as an example. Let $y_i(i=1,2,\cdots,m)$ be the impact values of the *i*-th KEGG pathway category and $x=(x_{i1},x_{i2},\cdots x_{ip})^T$ be the impact values of its corresponding subcategory. The vector x is assumed to follow a normal distribution, $x\sim N(0,R_x)$, where R_x is the correlation matrix of x. Let $y_i^{'}$ and $x_{ij}^{'}(j=1,2,\cdots,p)$ denote the standardized y_i and $x_{ij}(j=1,2,\cdots,p)$, the standardized multiple linear regression equations is

Molecular BioSystems Accepted Manuscript

$$y_i^{'} = \sum_{j=1}^{n} b_j^* x_{ij}^{'} + \varepsilon_i \qquad (1)$$

Where $y_i^{'} \sim N(0,1)$, $\varepsilon_i \sim N(0,1)$ and ε_i is independent. Using the least squares estimation method, we can easily get the canonical equations to solve the path coefficients as follows:

$$\begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_p^* \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \\ \vdots \\ r_{py} \end{bmatrix} \quad \text{Or} \quad \hat{R}_x b_j^* = \hat{R}_{xy} \qquad (2)$$

Where $\hat{R}_x$ is the maximum likelihood estimation of correlation matrix R_x. And $\hat{R}_{xy}$ is the correlation matrix of x and y_i, which is called the total effect reflecting the importance of each subcategory pathway to corresponding category pathway. For example, if the total effect of subcategory pathway x_{ij} is the largest, then this subcategory pathway is regarded as the most important pathway in all subcategory pathways of corresponding category pathway y_i. In fact, equations (2) have performed the subdivision of the total effect. The solved path coefficient result $b^* = (b_1^*, b_2^*, \cdots, b_p^*)^T$ indicates the direct effect of each subcategory pathway on y_i. In addition, $r_{jt} b_t^* \ (j = 1,2,\cdots,p; t = 1,2,\cdots,p; t \neq j)$ indicates the indirect effect on y_i (x_{ij} through x_{it}). The subdivided results can be showed as Table 1. Obviously, the detailed subdivided results can fully display the direct and indirect effect of specified subcategory pathway. The phenomenon that the direct effect of specified subcategory pathway is larger than the indirect effect from the other pathways indicated that this subcategory pathway was directly impacted to a large extent. Otherwise, this

Molecular BioSystems Accepted Manuscript

subcategory pathway mainly impacted by the indirect regulation from the other related subcategory pathways. This subdivision also can be demonstrated visually as Figure 2.

In short, these three parameters (b_j^* , $r_{jt}b_t^*$, r_{jy}) can illustrate the complex regulating interactions between KEGG pathway category and its corresponding subcategory pathways in detail and measure the importance of each subcategory pathway. Similarly, the interactions between KEGG pathway subcategory and its secondary pathways can be analyzed through this subdivision.

2.2 The gradient analysis method from principal component analysis (PCA)

We introduce the gradient analysis method from principal component analysis in KEGG pathway subcategories and the secondary pathways in order to estimate the impact direction of each pathway.

The same as above assumption, let $x = (x_{i1}, x_{i2}, \cdots x_{ip})^T$ be the impact values of the subcategory pathways belonging to the *i*-th KEGG pathway category. The vector x is assumed to follow a normal distribution, $x \sim N(0, R_x)$, where R_x is the correlation matrix of x. We use $\lambda_1 > \lambda_2 > \cdots \lambda_p > 0$ and $U_j = \left(U_{j1}, U_{j2}, \cdots, U_{jp}\right)^T$ to denote the eigenvalue of R_x and the corresponding eigenvector of λ_j, respectively. Then we have $F_j = U_j^T x$ and F_j are called as the *j*-th principal component. The variance of F_j is λ_j, that is $V(F_j) = \lambda_j$ and $\sum_{j=1}^{p} \lambda_j = p$. The first l principal components were chosen to estimate the impact direction of the pathway, meanwhile

Molecular BioSystems Accepted Manuscript

their accumulative variance contribution rate is greater than or equal to 99% ($\sum_{j=1}^{l}\left(\lambda_j \Big/ p\right) \geq 99\%$). The other remaining principal components will be discarded as the random noise. Obviously,

$$\left(\frac{\partial F_j}{\partial x_{i1}}, \frac{\partial F_j}{\partial x_{i2}}, \cdots, \frac{\partial F_j}{\partial x_{ip}}, \right) = \left(u_{j1}, u_{j2}, \cdots, u_{jp}\right),$$

so the principal component analysis can be regarded as the direct gradient analysis under the variances maximization. The result $\frac{\partial F_j}{\partial x_{ik}}$ is the gradient of the subcategory pathway x_{ik}. In fact, the sum $\sum_{j=1}^{l}\left(\frac{\partial F_j}{\partial x_{ik}}\right)$ reflects the dynamic variance of the KEGG pathway x_{ik}. Therefore, when $\sum_{j=1}^{l}\left(\frac{\partial F_j}{\partial x_{ik}}\right) \geq 0$, the impact direction of this pathway can be regarded as up-regulated, while $\sum_{j=1}^{l}\left(\frac{\partial F_j}{\partial x_{ik}}\right) \leq 0$, the impact direction is mainly down-regulated. Similarly, the gradient analysis method can be used in the secondary pathways.

3. Application

3.1 Dataset

In order to test the utility of the KEGG-PATH approach, we selected the KEGG pathway DIA impact values from the time-course experiment--functional analysis of the bovine mammary transcriptome during the lactation cycle. The detailed impact values data are attached in Table S1. The 'Human Diseases' category-related pathways have been discarded from our analysis in that it almost has nothing to do

Molecular BioSystems Accepted Manuscript

with bovine mammary according to the document [6]. The real data from the KEGG pathways are processed as follows: firstly, to uncover the dynamic nature of the regulation mechanisms among KEGG pathways during entire lactation cycle, we chose the impact values data only from -15 to 300 vs. -30d. Secondly, the pathways were deleted when the number of its corresponding missing data was greater than or equal to three. Finally, if the number of the missing data included in the pathway is less than three, we filled it with the average value of the other values belonging to this pathway. The filled data were marked in red color in Table S1.

In addition, in order to compare the results of impact direction produced by the gradient analysis method from PCA and the DIA method, the impact direction data of pathways calculated by DIA method from -15 to 300 vs. -30d were listed in Table S2. The missing data were filled by the average value of the other values belonging to this pathway. And the filled data were also marked in red color.

3.2 Results

The overall results of the KEGG pathway categories and its subcategories, the subcategories and its secondary pathways based on the KEGG-PATH approach were shown in Table S3, which denoted the detailed total effect subdivision. The detailed comparisons between the total effect from KEGG-PATH approach and average impact values from DIA method, and the impact direction results produced by the DIA method and the gradient analysis from PCA were listed in Table S4.

3.2.1 The results of KEGG-PATH approach

Molecular BioSystems Accepted Manuscript

Overall interpretation of pathways: Metabolism In the category 'Metabolism', as Table S3 (a) sheet showed that the direct effect of subcategory 'Glycan Biosynthesis and Metabolism' was greater than the indirect effect from all the other pathways, which strengthened the total effect. The result indicated that this subcategory was activated directly. The direct effect of 'Lipid Metabolism', 'Nucleotide Metabolism' and 'Amino Acid Metabolism' was similar with their indirect effect with each other, but was greater than the indirect effect from the other pathways. The result showed the impact effect of these pathways integrated the direct effect with the indirect effect. Hence, the regulation of these pathways was as important as their direct effect. It is worth noting that these pathways were all up-regulated to a large extent by the pathway 'Glycan Biosynthesis and Metabolism'. These results were expected, especially the importance of pathways related to 'Lipid Metabolism', given that the substantial amount of milk fat was produced by the mammary gland. The fact that 'Amino Acid Metabolism' and 'Nucleotide Metabolism' pathways were impacted largely during lactation appeared to support the importance of milk protein synthesis in bovine mammary. The total effect of remaining pathways was mainly controlled by the correlation regulation.

In detail, the results (Table S3 (b)) demonstrated that in subcategory 'Carbohydrate Metabolism', although the total effect of pathway 'Citrate cycle (TCA cycle)' is the largest, the direct effect is very small. The result showed that 'TCA cycle' activity was mainly regulated by the other pathways. On the contrary, the total

Molecular BioSystems Accepted Manuscript

effect of pathway ‘Pyruvate metabolism’ is large, and its direct effect is far greater than the indirect effect from the other pathways. The result indicated that the ‘Pyruvate metabolism’ was highly impacted directly. In addition, the direct effect of important pathway ‘Galactose metabolism’ related to lactose synthesis was placed in the middle of all the direct and indirect effects. The result demonstrated that the activation of the pathway ‘Galactose metabolism’ depended not only on its direct impaction, but also on the correlation regulation of the other pathways.

In ‘Lipid Metabolism’ subcategory, the direct effect of pathway ‘Arachidonic acid metabolism’ was prominent, which showed this pathway was impacted directly. Meanwhile, the pathways ‘Glycerolipid metabolism’, ‘Fatty acid metabolism’ and ‘Sphingolipid metabolism’ had the relatively greater direct suppressed effect and the negative indirect regulation on the other pathways. Next, it seems that the impaction of the pathways ‘Ether lipid metabolism’ and ‘Steroid biosynthesis’ from the direct and indirect effect was alike. By contrast, although the total effect of pathway ‘Glycerophospholipid metabolism’ was large, the direct effect was very small. The result showed that this pathway was impacted mainly by the correlation regulation. To our surprise, the total effect of ‘Biosynthesis of unsaturated fatty acids’ was not apparently large. The result was unexpected because this pathway was very important for the milk fat synthesis.

In the subcategory ‘Amino Acid Metabolism’, the direct effect of pathway ‘Lysine degradation’ was far greater than the indirect correlation regulation from the

Molecular BioSystems Accepted Manuscript

other pathways. Obviously, the pathway 'Lysine degradation' was impacted directly. In addition, the direct effect of pathways 'Arginine and proline metabolism', 'Histidine metabolism' and 'Valine, leucine and isoleucine biosynthesis' was less than the indirect effect of 'Lysine degradation', but greater than the indirect effect from the other pathways. The result demonstrated that only the negative regulation of 'Lysine degradation' exceeded the direct effect of these pathways. However, the direct effect of 'Tryptophan metabolism' and 'Valine, leucine and isoleucine degradation' was relatively smaller, which indicated that we should attach importance to the correlation regulation from the other pathways.

The importance of subcategory 'Glycan Biosynthesis and Metabolism' in mammary during lactation was evidenced mainly by the pathways 'N-Glycan biosynthesis', 'O-Mannosyl glycan biosynthesis', 'Glycosphingolipid biosynthesis - ganglio series' and 'Glycosylphosphatidylinositol (GPI)-anchor biosynthesis'. And glycosphingolipid biosynthesis is important because of its membrane function and of particular interest due to its current commercial focus [7]. Among of them, only the direct effect of 'N-Glycan biosynthesis' pathway was small, the direct effect of the other pathways was relatively greater. The result showed that pathway 'N-Glycan biosynthesis' was impacted mainly through the regulation from the other pathways; conversely, the other three pathways were impacted directly to a great degree.

Overall interpretation of pathways: Genetic Information Processing As Table S3 (a) showed that the direct effect of 'Folding, Sorting and Degradation'

Molecular BioSystems Accepted Manuscript

subcategory pathway was slightly greater than the pathway 'Translation' and 'Replication and Repair' and it had relatively large up-regulating effect on the other pathways. Among the KEGG secondary pathways of subcategory 'Folding, Sorting and Degradation', the direct effect of pathway 'Ubiquitin mediated proteolysis' and 'SNARE interactions in vesicular transport' was obviously far greater than the indirect effect from the other pathways. The result indicated that these two pathways were impacted directly. In addition, the total effect of the pathway 'Protein processing in endoplasmic reticulum' and 'RNA degradation' was large, but their direct effect was very small. Obviously, these two pathways were regulated mainly by the other pathways. The total effect of pathway 'Ribosome biogenesis in eukaryotes', within subcategory 'Translation', was larger, which showed that this pathway was largely impacted. The result seems to support the apparent increase in protein synthesis inferred by transcriptome analysis suggested by Finucane et al [8]. In subdivision, its direct effect and indirect effect were almost equivalent. The phenomenon demonstrated that the impaction of this pathway was the joint result of the direct effect and the regulation effect. This result was dissimilar to the pathway 'mRNA surveillance pathway', which not only had the larger total effect, but also had the larger direct effect. The relatively small indirect effect indicated that it was directly induced. Differently, the direct effect of pathway 'Ribosome' was relatively large, but its total effect was small. In the secondary pathways subdivision of subcategory 'Replication and Repair', the pathways 'Homologous recombination' and 'DNA

Molecular BioSystems Accepted Manuscript

replication' had the larger total effect and direct effect, which indicated they were induced directly. But the pathway 'Nucleotide excision repair' was impacted by the regulation from the other pathways.

Overall interpretation of pathways: Environmental Information Processing In the subcategory-subdivision of this category pathway, the direct effect of 'Signal Transduction' pathway was obviously greater than the indirect effects from the other pathways and it was a positive regulation factor for the other pathways. This result emphasized its importance in the whole category pathways. More importantly, all its secondary pathways were impacted largely. And in the subdivision of its secondary pathways, the direct effects of pathways 'Notch signaling pathway' and 'Jak-STAT signaling pathway' and their indirect effects with the other pathways were positive and relatively larger, which demonstrated that the two pathways were highly activated and they had the large up-regulating effect on the other pathways. The Jak-STAT signaling has been previously reported to be essential for the induction of milk protein expression in mammary tissue of non-ruminants [9], but is not clear in bovine. The activation of Notch signaling pathway was unexpected given that in mammary cell lines it has been observed that artificial activation of Notch signaling inhibits lactation[10]. On the contrary, the direct effects of pathways 'Calcium signaling pathway' and 'ErbB signaling pathway' and their indirect effects with the other pathways were negative and relatively larger. This result showed that these two pathways were highly inhibited and had the large down–regulating effect on the other

Molecular BioSystems Accepted Manuscript

pathways. The inhibition of the Calcium signaling was a novel finding. It has been reported that superfluous calcium will lead to the inhibition of lactation [11]. It is worth noting that the obvious activation of pathway 'Wnt signaling pathway' was in accordance with the previous fact that Wnt signaling appear to have a role in mammary stem cell self-refresh [12]. In addition, the total effect of pathway 'mTOR signaling pathway' was relatively larger. The result was expected given that mTOR seems to play a role in milk protein synthesis. But its direct effect was far less than the indirect effects from the other pathways, which demonstrated that mTOR signaling was not inhibited directly.

Overall interpretation of pathways: Cellular Process. In the path chain of subcategory 'Transport and Catabolism', as Table S3 (a) sheet showed that the direct effect was the largest and the indirect effect from the pathways 'Cell Communication' and 'Cell Growth and Death' succeeded. The indirect effect of pathway 'Cell Motility' was relatively small. The result illustrated that 'Transport and Catabolism' pathway was highly activated directly and it received the positive regulation from 'Cell Communication' and 'Cell Growth and Death'. But the regulation of 'Cell Motility' was smaller. However, the direct effect of 'Cell Motility' was far less than the indirect effect from the other pathways, which showed that the activation of pathway 'Cell Motility' was mainly due to the up-regulating effect from the other three pathways. The results of subcategory 'Cell Growth and Death' and 'Cell Communication' path chain illuminated that the indirect effect of 'Transport and

Molecular BioSystems Accepted Manuscript

Catabolism' was greater than their own direct effects, which gave more prominence to the up-regulated effect of the pathway 'Transport and Catabolism'. In short, in category 'Cellular Process', the subcategory pathway 'Transport and Catabolism' was the most activated and it regulated largely the other pathways. The pathway 'Cell Growth and Death' was highly impacted, which appeared to be in contradiction with the reduction of proliferation and stabilization of cell numbers during most of lactation[8, 13-15].

In Table S3 (b) sheet, the direct effect of the secondary pathways 'Peroxisome' and 'Phagosome' belonging to subcategory 'Transport and Catabolism' was all obviously large. At the same time, they had the larger up-regulated effect on the other pathways. It was worthwhile to note that pathway 'Endocytosis' was suppressed slightly and it had the down-regulating effect on the other pathways. In addition, the direct effects and indirect regulation of four secondary pathways from subcategory 'Cell Growth and Death' were almost equivalent from the point of subdivision. In subcategory 'Cell Communication', the secondary pathways 'Tight junction' and 'Gap junction' had the relatively larger direct effects and positive correlated regulation effects. This result can be proved by the fact that the junctions in mammary tissue were important during lactation, particularly the 'Tight junction' [16-17].

Overall interpretation of pathways: Organismal Systems The 'Organismal Systems' category of KEGG pathways was highly impacted, with the exception of pathway 'Environmental Adaptation'. However, due to specific functions were not

Molecular BioSystems Accepted Manuscript

pertinent to the mammary, only the subcategory pathways 'Immune System', 'Endocrine System' and 'Nervous System' were analyzed in detail. As Table S3 (a) sheet showed that the direct effect of pathway 'Endocrine System' was far greater than the indirect effect from the other pathways, which indicated that this pathway was impacted directly to a great degree. Among the secondary pathways composing the 'Endocrine System', the direct effect of 'Insulin signaling' and 'GnRH signaling' pathways was positive and the largest, which illustrated that these two pathways were highly activated directly, although they were down-regulated by some other pathways through the negative correlation. An increase in expression of genes on GnRH signaling in mammary tissue during lactation had been reported in mouse [18-19], but to our knowledge, no report was in bovine. The activation of pathway 'Insulin signaling' happened to agree with the slightly inhibition of mTOR signaling given that the inhibition of the mTOR signaling would be overridden by insulin signaling [6]. On the contrary, the direct effect of pathway 'Adipocytokine signaling pathway' was far less than the indirect effects from the other pathways, which indicated that this pathway was impacted mainly by the regulation. However, in the path chains of subcategory pathways 'Immune System' and 'Nervous System', the direct effect was obviously far less than the indirect effect. The result showed that these two pathways were impacted mainly due to the correlation regulation. Furthermore, the detailed subdivision of the secondary pathways illustrated that the activation of 'Immune System' could be characterized by the obvious direct up-regulated pathways 'Toll-like receptor

Molecular BioSystems Accepted Manuscript

signaling pathway', 'Leukocyte transendothelial migration', 'Hematopoietic cell lineage' and 'B cell receptor signaling pathway' , which were well-known related to the innate immune system and immune cells, and the down-regulated pathways 'Intestinal immune network for IgA production', 'Chemokine signaling pathway' and 'Fc epsilon RI signaling pathway'. In addition, the direct and indirect effects of four secondary pathways from subcategory 'Nervous System' were almost equivalent from the point of subdivision.

3.2.2 The results of comparisons

From the perspective of the impact importance comparison, the KEGG-PATH method can produce more biologically meaningful results. Firstly, almost all highly impacted pathways captured by DIA method were also found by the KEGG-PATH approach. Secondly, KEGG-PATH approach can find some other more biologically meaningful results. For example, the total effect of subcategory 'Amino Acid Metabolism' ranked second in metabolism category, which can illustrate the biological phenomenon of 'milk protein synthesis', while this result was not captured by DIA method. In addition, in category 'Environmental Information Processing', the total effect of subcategory 'Signal Transduction' unexpectedly ranked first, which was quite opposite to the DIA result. But it can be supported by the fact that almost all its secondary pathways were highly activated. And the order of the total effect on subcategory 'Nervous System'（belonging to the KEGG category 'Organismal Systems'） was brought forward to the fourth. It was well-known that the nervous

Molecular BioSystems Accepted Manuscript

system in mammary gland during lactation was important due to its role in milk ejection [20] and its likely control of blood flow [21].

The main results of impact direction comparison between gradient analysis method from principal component analysis and DIA method were just as follows. Firstly, as Table 2 showed that the concordance rate of impact direction was from 33.3% to 88.9% from the point of each category pathway. Secondly, from the view of the whole subcategory pathways, the concordance rate was 64.5%; but for the whole secondary pathways, the concordance rate was 48.8%. In fact, for the secondary pathways of specific subcategory, the concordance rate was even up to 100%. For example, the directions of all secondary pathways of subcategory 'Nervous System' and 'Xenobiotics Biodegradation and Metabolism' were in complete agreement from these two methods.

4. Discussion

In this article, we proposed KEGG-PATH, a path analysis approach for KEGG-based pathway analysis. The model has a relatively simple format, with the nature as a standard multiple linear regression model and a special case of structural equation modeling. The key innovation is that it can subdivide the total effect deeply between KEGG pathway categories and subcategories, subcategories and the secondary pathways (Table S3). In this way, the complex regulating mechanisms among the KEGG pathways can be demonstrated clearly. In addition, we have given detailed comparisons between the total effect from KEGG-PATH approach and the

Molecular BioSystems Accepted Manuscript

average impact values from DIA method, and the impact direction results produced by the DIA method and the gradient analysis (Table S4). The comparisons suggest that the KEGG-PATH approach can produce more biologically meaningful results than the DIA method. The concordance rate of impact direction estimation for gradient method from PCA and DIA method was relatively high. Hence, the KEGG-PATH approach is a kind of data mining of functional enrichment analysis.

For the confirmation of the most impacted pathways, the DIA method only averages the impact values of the pathway during different time course, which does not consider its interaction effects with the other related pathways. The KEGG-PATH approach, on the other hand, borrows the total effect to judge the importance of the pathways, which needs not only to consider the direct effect itself but also to add the correlation indirect effect from the other related pathways. For the estimation of the impact direction, the DIA method averages the impact direction values of the pathway during different time course in the same way; but the gradient analysis method from PCA was a statistics calculation under the dimensionality reduction and the elimination of random interference. We have developed a program in Matlab (R2008a, version 7.6.0.324) to implement the KEGG-PATH approach and the gradient analysis method from PCA (S1). In the calculations, it might also be noted that the results seldom may be inaccurate when the correlation matrix is close to singular or badly scaled. But the relative error is basically controlled to 10^{-15} and it can be neglected.

Molecular BioSystems Accepted Manuscript

Currently, the proposed method is based on the DIA impact values. We are working on generalizing the KEGG-PATH approach to cases with the original gene expression value. Although we have focused on the use of this approach for the interpretation of the KEGG pathways, the general strategy can be applied to any circumstance in which groups of entities are annotated with similar dependency structure.

5. Conclusions

Overall, our analysis indicated that KEGG-PATH approach can deeply uncover the complex regulation relationship of KEGG pathways through the subdivision of the total effect. This approach is a kind of data mining through functional enrichment analysis of time-course experiments or those involving multiple treatments.

Molecular BioSystems Accepted Manuscript

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (Grant Nos. 31372279).

References:

[1] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, R. Romero, *Genome Res.,* 2007, 17:1537-1545. [PubMed: 17785539]

[2] S. Zhang, J. Cao, Y. M. Kong, R. H. Scheuermann, *Bioinformatics*, 2010, 26[7]: 905-911. [PubMed: 20176581]

[3] W. Huang da, B. T. Sherman, R. A. Lempicki, *Nucleic Acids Res.*, 2009, 37(1): 1-13. [PubMed: 19033363]

[4] M. Bionaz, K. Periasamy, S. L. Rodriguez-Zas, W. L. Hurley, J. J. Loor, *Plos one*, 2012, 7(3): e32455. [PubMed: 22438877]

[5] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, *Nucleic Acids Res*, 1999, 27:29-34. [PubMed: 9847135]

[6] M. Bionaz, K. Periasamy, S. L. Rodriguez-Zas, R. E. Everts, H. A. Lewin, W. L. Hurley, J. J. Loor, *Plos one*, 2012, 7(3): e33268. [PubMed: 22428004]

[7] N. Tao, E. J. DePeters, S. Freeman, J. B. German, R. Grimm, C. B. Lebrilla, *J. Dairy Sci.*, 2008, 91: 3768-3778. [PubMed: 18832198]

[8] K. A. Finucane, T. B. McFadden, J. P. Bond, J. J. Kennelly, F. Q. Zhao, *Funct. Integr. Genomics*, 2008, 8: 251-264. [PubMed: 18259788]

[9] M. Bionaz, J. J. Loor, *Bioinform. Biol. Insights*, 2011, 5: 83-98. [PubMed: 21698073]

Molecular BioSystems Accepted Manuscript

[10] R. Callahan, S. E. Egan, *J. Mammary Gland Biol.*, 2004, 9: 145-163. [PubMed: 15300010]

[11] H. Chu, Z. Wang, F. Li, *Chin. J. Anim. Nutrition*, 2010, 22(5): 1286-1292.

[12] W. A. Woodward, M. S. Chen, F. Behbod, J. M. Rosen, *J. Cell Sci.*, 2005, 118: 3585-3594. [PubMed: 16105882]

[13] J. V. Norgaard, P. K. Theil, M. T. Sorensen, K. Sejrsen, *J. Dairy Sci.*, 2008, 91: 2319-2327. [PubMed: 18487654]

[14] A. V. Capuco, D. L. Wood, R. Baldwin, K. McLeod, M. J. Paape, *J. Dairy Sci*., 2001, 84: 2177-2187. [PubMed: 11699449]

[15] A. V. Capuco, S. E. Ellis, S. A. Hale, E. Long, R. A. Erdman, X. Zhao, M. J. Paape, *J. Anim. Sci*., 2003, 81 Suppl. 3: 18-31. [PubMed: 15000403]

[16] D. R. Pitelka, S. T. Hamamoto, *Ultrastructure of the Mammary Secretory Cell. In "Biochemistry of Lactation",* (ed: Mepham T. B.), Amsterdam: Elsevier Science Publishers B.V., 1983, pp 29-70.

[17] D. A. Nguyen, M. C. Neville, *J. Mammary Gland Biol.*, 1998, 3: 233-246. [PubMed: 10819511]

[18] M. C. Rudolph, J. L. McManaman, L. Hunter, T. Phang, M. C. Neville, *J. Mammary Gland Biol.*, 2003, 8: 287-307. [PubMed: 14973374]

[19] R. W. Clarkson, M. T. Wayland, J. Lee, T. Freeman, C. J. Watson, *Breast Cancer Res*., 2004, 6: R92-109. [PubMed: 14979921]

[20] C. E. Grosvenor, F. Mena, *Neural and Hormonal, Control of Milk Secretion and Milk*

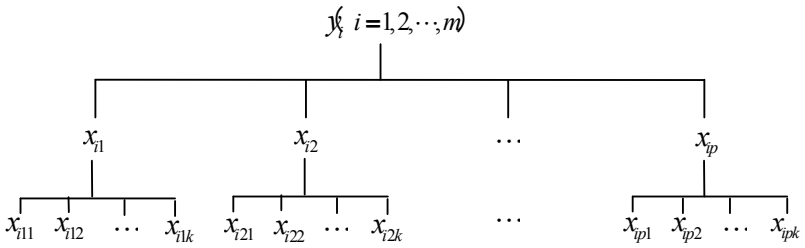
Molecular BioSystems Accepted Manuscript

Ejection. In "Lactation: a comprehensive treatise" (eds.: B. L. Larson, V. R. Smith), New York: Academic Press, 1974, pp 227-276.

[21] J. L. Linzell, *Mammary Blood Flow and Substrate Uptake. In "Lactation: a comprehensive treatise"* (eds.: B. L. Larson, V. R. Smith), New York: Academic Press, 1974, pp 143-225.

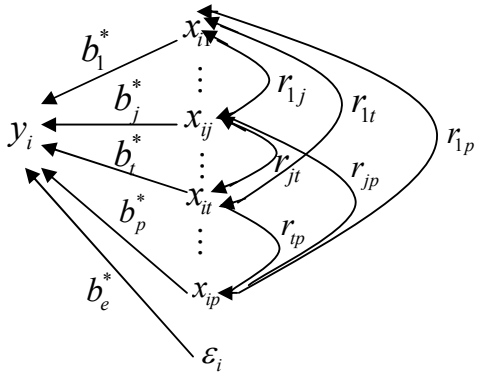
Molecular BioSystems Accepted Manuscript

F**igure 1:** the network data structure of the KEGG pathway



Molecular BioSystems Accepted Manuscript

Figure 2: the completely closed path chart with independent error ε_i



Molecular BioSystems Accepted Manuscript

Table 1: The detailed subdivided result of subcategory pathway

Subcategory pathway/the secondary pathway	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{it}	$\cdots$	x_{ip}
the direct effect (b_j^*) and indirect effect ($r_{jt}b_t^*$) ($j=1,2,\cdots,p;\quad t=1,2,\cdots,p;\quad j\neq t$)	b_1^*	$\cdots$	$r_{j1}b_1^*$	$\cdots$	$r_{t1}b_1^*$	$\cdots$	$r_{p1}b_1^*$
	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$r_{1j}b_j^*$	$\cdots$	b_j^*	$\cdots$	$r_{tj}b_j^*$	$\cdots$	$r_{pj}b_j^*$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
	$r_{1t}b_t^*$	$\cdots$	$r_{jt}b_t^*$	$\cdots$	b_t^*	$\cdots$	$r_{pt}b_t^*$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$r_{1p}b_p^*$	$\cdots$	$r_{jp}b_p^*$	$\cdots$	$r_{tp}b_p^*$	$\cdots$	b_p^*
The total effect (r_{jy})	r_{1y}	$\cdots$	r_{jy}	$\cdots$	r_{ty}	$\cdots$	r_{py}

Note: $r_{jt}\ (j=1,2,\cdots,p;t=1,2,\cdots,p;j\neq t)$ indicates the correlation coefficient x_{ij} and x_{it}. Obviously, the data satisfy $r_{jt}=r_{tj}$ and $r_{jy}=b_j^*+\sum_{\substack{t=1\\t\neq j}}^{p} r_{jt}b_t^*$ according to the Path analysis method. In order to distinguish between the direct and indirect effect clearly, the direct effect has been marked using the red frame.

Molecular BioSystems Accepted Manuscript

Table 2: The comparison of impact direction between gradient methods from PCA with DIA method

Category	Concordance rate	
	in its subcategory pathways	**in its secondary pathways**
Metabolism	45.5% (5/11)	49.3% (34/69)
Genetic Information Processing	75% (3/4)	47.6% (10/21)
Environmental Information Processing	33.3% (1/3)	57.1% (8/14)
Cellular Processes	75% (3/4)	38.5% (5/13)
Organismal Systems	88.9% (8/9)	48.8% (21/43)
sum	64.5% (20/31)	48.8% (78/160)

Note: For the first five lines, in the parentheses of "**in its subcategory pathways**" column, the denominator of each fraction denotes the number of subcategory pathways from the front corresponding category, and the numerator of each fraction denotes the number of pathways with the same impact direction under these two methods; in the parentheses of "**in its secondary pathways**" column, the denominator of each fraction denotes the number of all secondary pathways from the front corresponding category, and the numerator of each fraction denotes the number of pathways with the same impact direction under these two methods.

Molecular BioSystems Accepted Manuscript