GO based Tissue Specific Functions of Mouse using Countable Gene Expression Profiles

YOICHI TAKENAKA[1] AKIKO MATSUMOTO[2] HIDEO MATSUDA[1]
takenaka@ist.osaka-u.ac.jp matsuda@ist.osaka-u.ac.jp

[1] *Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University Machikaneyama 1-3, Toyonaka City, Osaka 560-8531, Japan*
[2] *JASTEC Co., Ltd.,Takanawa 3-5-23, Minato-ku, Tokyo Japan*

We present a new method to describe tissue-specific function that leverages the advantage of the Cap Analysis of Gene Expression (CAGE) data. The CAGE expression data represent the number of mRNAs of each gene in a sample. The feature enables us to compare or add the expression amount of genes in the sample. As usual methods compared the gene expression values among tissues for each gene respectively and ruled out to compare them among genes, they have not exploited the feature to reveal tissue specificity. To utilize the feature, we used Gene Ontology terms (GO-terms) as unit to sum up the expression values and described specificities of tissues by them. We regard GO-terms as events that occur in the tissue according to probabilities that are defined by means of the CAGE. Our method is applied to mouse CAGE data on 22 tissues. Among them, we show the results of molecular functions and cellular components on liver. We also show the most expressed genes in liver to compare with our method. The results agree well with well-known specific functions such as amino acid metabolisms of liver. Moreover, the difference of inter-cellular junction among liver, lung, heart, muscle and prostate gland are apparently observed. The results of our method provide researchers a clue to the further research of the tissue roles and the deeper functions of the tissue-specific genes. All the results and supplementary materials are available via our web site.

Keywords: Tissue Specificity; Gene Ontology; Expression Profile; CAGE

1. Background

Analysis of gene expression has contributed to reveal the roles and functions of genes. Many researchers have studied genes with similar expression pattern [1] and have inferred the regulation networks of genes [2, 3]. The gene expressions are also used to reveal the roles and functions of tissues from the view point of genes. It is reasonable to infer that over-expressed genes in a tissue closely related to the functions of the tissue, and the genes that are not highly but expressed only in a particular tissue are also be related to. Hereafter, we call these tissue-specific genes. To observe gene expression, DNA microarrays [1, 4], EST (Expression Sequence Tag) [5], SAGE (Serial Analysis of Gene Expression) [6, 7], MPSS (Massively Parallel Signature Sequencing) [8], and CAGE (Cap Analysis of Gene Expression) [9, 10] has been widely used. Using these experimental data, Schug et al. found tissue-

specific promoter features as measured by Shannon entropy [11], and Kadota et al. proposed a method to identify tissue-specific genes by outlier detection [12, 13]. When genes turn out to be tissue-specific, the functions of the genes help us to reveal the tissue-specific functions.

However, functions of the tissue-specific genes are insufficient to describe tissue-specific functions. Firstly, the genes that express significantly in a tissue not necessarily mean that the genes express higher than other genes. Of course, the tissue-specific genes with lower expression amount will be influenced to the inherent function of the tissue, but they can be lightly affected to than tissue-specific genes with higher expression amount. Secondly, it is natural that a set of genes contribute a role of tissues. For examples, hemoglobin is a hetero-dimer protein one of whose molecular function is oxygen binding. Biological process like chitin metabolisms and cellular components such as desmosome or connexson complex are closely-related to the functions of the tissue. As usual methods only focused on functions of genes, they wrote off the functions composed by a set of the genes. The aim of our study is to propose a method to determine, as completely as possible, tissue-specific functions that are hard to be identified from the tissue-specific genes described by usual methods.

To describe tissue-specific functions, we use Gene Ontology (GO) [14]. GO provides a set of structured vocabularies for specific biological domains that can be used to describe the domain in terms of their associated biological process, cellular components and molecular functions, in a species-independent manner. Each word is called a GO-term, and relationships among GO-terms are described as a directed acyclic graph, such that the child GO-term has more specific meanings than the parent GO-term. At present, GO is the only solution to describe the function of genes systematically and computationally and has become de facto standard for gene annotations [15, 16].

To measure tissue specificity of the functions by gene expression profiles, it is required to sum up the expression value of the genes along the hierarchical structure of gene ontology. Therefore, the expression values are required to be addable. The expression profiles also need to be measured cyclopedically. Among the gene expression-measuring methods, CAGE is the only one that satisfies these two conditions [17]. CAGE is a sequence based technology with which to collect 20bp tags from the 5'-end of transcripts in a sample exhaustively [9, 10, 17–19]. The collected tags are sequenced and mapped to the genome to identify which genes are expressed. As the number of the mapped tags represents the number of mRNA of the target gene in the sample, the gene expression data measured by CAGE is addable.

EST and SAGE also measure gene expression by sequencing transcripts. EST makes cDNA clones from transcripts and determines clone sequences from the 5'-end or 3'-end. The sequence lengths are several hundred bases. EST has contributed greatly to the detection of genes, gene annotations and establishment of the gene expression profiles. However, the sequences used as EST are too long to identify the expressed gene on a cost. Therefore, EST has been replaced with a more efficient

method, namely SAGE. SAGE collects 10bp tags from the 3'-end of transcripts. As 10 to 20 tags are concatenated at a time to determine the sequence, SAGE is an efficient method to observe gene expression. However, SAGE requires the recognition sequence of NlaIII, namely CATG, near the 3'-end of the transcripts in order to collect the tags. It disturbs SAGE to observe the expression of genes exhaustively.

DNA microarrays and MPSS use competitive hybridization to observe gene expression. They anchor complementary strands of target genes on glass plates or solid beads, dye genes from the target condition and control condition with different colors, and then competitively hybridize the genes to the complementary strands. Gene expression is measured as the proportion of the brightness of the two colors. These methods measure the gene expression in the target condition based on the magnification factor over the control. Therefore, they cannot compare the expressions of two genes in the same condition. More directly saying, it is impossible to know which gene is more highly expressed. The gene expression data from these methods is not addable, and cannot be used to determine tissue-specific functions.

One of the simplest ways to measure the tissue-specificity of a function is the summation of the expression amount of the genes that play a role of the function. It can be reasonable, but includes an imperfection. The functions of house-keeping genes, which are expressed highly in all tissues and consequently they are inappropriate as tissue-specific, will be over estimated. To avoid the imperfection, we applied the information content that is an idea in the field of information theory [20] to our measurement of tissue-specificity.

Let E be an event that occurs according to the probability p, then information content of the event E is defined as $-\log(p)$. In this paper, the events correspond to the functions or GO-terms. The probability of a GO-term is defined as X/Y, where X is the summation of the CAGE tags of the genes that have the GO-term and Y is the summation of all the tags in the tissue. The information content of the GO-term is $-\log(X/Y)$. We defined the tissue-specificity of a GO-term as the difference in information content between the target tissue and whole tissues. We used 11,567,973 CAGE tags from 22 tissues of mouse offered by the FANTOM consortium [17] for the analysis of tissue-specific functions.

2. Results

Twenty two tissues of mouse CAGE data [17] were used to calculate tissue specific functions, and we selected liver, in which the largest number of tags are measured among 22 tissues, to show the results. GO-terms had three main categories: molecular function, biological process and cellular component. Due to space limitation, we chose molecular function and cellular component to show the result. As no usual method were utilize the countable expression data and the hierarchy of Gene Ontology, we show the Top 20 GO-terms of the most expressed genes for the comparison. The m̈ost expressed genesm̈eans that the numbers of mRNA in the

Table 1. Top20 TSFs of liver on Molecular Function

rank TSF	rank MEG	GO-term	EC number	#tag	#norm. tag
1	9139	biphenyl-2,3-diol 1,2-dioxygenase activity	1.13.11.39	3	0.9
2	4527	antifungal peptide activity		33	9.8
3	1672	endogenous peptide antigen binding		153	45.5
4	199	alcohol sulfotransferase activity	2.8.2.2	1214	361.3
5	120	tyrosine transaminase activity	2.6.1.5	1837	546.7
6	266	urocanate hydratase activity	4.2.1.49	941	280.0
7	140	betaine-homocysteine S-methyltransferase activity	2.1.1.5	2660	791.6
8	140	homocysteine S-methyltransferase activity	2.1.1.10	3030	901.7
9	103	urate oxidase activity	1.7.3.3	2219	660.3
9	-	oxidoreductase activity, acting on other nitrogenous compounds as donors		2219	660.3
9	-	oxidoreductase activity, acting on other nitrogenous compounds as donors, oxygen as acceptor		2219	660.3
12	14	4-hydroxyphenylpyruvate dioxygenase activity	1.13.11.27	7919	2356.6
12	14	quercetin 2,3-dioxygenase activity	1.13.11.24	7919	2356.6
14	369	plasmin activity	3.4.21.7	701	208.6
15	538	carbamoyl-phosphate synthase (ammonia) activity	6.3.4.16	513	152.7
15	538	carbamoyl-phosphate synthase (glutamine-hydrolyzing) activity	6.3.5.5	513	152.7
17	558	protein C (activated) activity	3.4.21.69	496	147.6
18	249	acyl-CoA ligase activity		992	295.2
18	249	butyrate-CoA ligase activity	6.2.1.2	992	295.2
20	275	triglyceride binding		907	269.9

cell are the largest among all the genes, which was a basic method that utilizes the countability to reveal the tissue specificity. The rankings of all the 22 tissues and the three GO-term categories are available from the certificated WEB page: *http://tsf.ics.es.osaka-u.ac.jp* with *(user,password) = (TSF,storia1441)*[a]. The supplementary tables are also placed the WEB page described above.

[a]Certification will be removed after the paper is accepted

Table 2. Top20 MEGs of liver on Molecular Function

rank TSF	rank MEG	GO-term	EC number	#tag	#norm. tag
241	1	ferric iron binding		62181	18504.3
1136	2	chaperone activity		43354	12901.6
1198	2	GTPase activity	3.6.5.1-4	61246	18226.0
1342	2	structural molecule activity		93385	27790.2
1218	2	structural constituent of cytoskeleton		53137	15812.9
1089	2	GTP binding		88836	26436.5
1129	2	MHC class I protein binding		43004	12797.4
736	3	peroxidase activity	1.11.1.7	30531	9085.6
376	3	glutathione peroxidase activity	1.11.1.9	20721	6166.3
407	3	oxidoreductase activity		265408	78982.1
382	4	MHC class I receptor activity		24287	7227.5
170	5	monooxygenase activity		40237	11974.0
71	5	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen		26818	7980.7
649	6	transcription corepressor activity		16399	4880.1
1237	6	protein binding		291171	86648.8
454	7	carrier activity		115575	34393.7
505	7	lipid binding		49978	14872.8
166	8	electron carrier activity		21581	6422.2
162	8	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen		22954	6830.8
108	8	cysteine dioxygenase activity	1.13.11.20	10529	3133.3

2.1. *Molecular Functions*

Table 1 shows top 20 GO-terms of molecular function calculated by our method (TSF), and table 2 shows top 20 GO-terms of the most expressed genes (MEG). The last rank of TSF was 2173rd and the last rank of MEG was 10311st. The last rank of TSF was equal to the number of GO-terms that was appeared in liver, and the last rank of MEG meant the number of genes that expressed in liver. In the tables, the column named TSF and MEG represent the rank of GO-term that

is shown in the column GO-term. When GO-terms are associated with Enzyme Commission number, we wrote them in the column EC number. The column named #tag shows the number of tags of the GO-term, which is calculated in **Step 2.** of our method. As the number of measured tags is different for each tissue, the column named #norm.tag shows the normalized number of tags by using a unit tpm (tags per million tags). There are "-" in the column of MEG rank. It represents that none of the genes with the GO-term expressed. The obvious differences between the two tables are #tag and the number of enzymes. The average #tag of TSF and MEG are 1874 and 69067, respectively. It represents that GO-terms in the top 20 TSF are not selected merely by the number of tags. The major part of the top 20 MEG rank was occupied by the common or abstract GO-terms, such as chaperone activity, GTPase activity, and structural molecule activity. They appeared in the MEG ranks of almost all tissues, and accordingly had lower TSF ranks. For example, ferric iron binding was ranked at 1st MEG, because *Trf*, whose gene description is transferin, was the most expressed gene in the liver. It also held 1st MEG rank at hippocampus, somatosensory cortex and visual cortex, and held 2nd to 10th MEG rank at 10 tissues. It was a typical function of the house keeping genes. There are 14 enzymes in the the TSF rank and four enzymes in the MEG rank. The frequent appearance of the enzymes in the top 20 TSF rank resulted from the intermediate metabolism that is one of the main roles of liver. Table 3 shows the number of enzymes that appeared in the Top20 TSF and MEG rank for each tissue. The average number of enzymes in the TSF rank was 5.3 and the average in the MEG rank was 2.9. In the table, liver and prostate grand had 14 enzymes in the TSF rank, which is the largest number among the 22 tissues, and followed by embryo and heart with 7 enzymes. To clarify the difference between liver and prostate grand, we categorized the enzymes according to the metabolic pathways in the KEGG (Kyoto Encyclopedia of Genes and Genomes) [21]. Nine of the 14 enzymes in the liver are responsible for metabolism, and of these, seven enzymes were involved in amino acid metabolism[b]. On the other hand, the prostate gland was rich in glycolipid metabolism. Five of the 14 enzymes were for glycolipid metabolism, two for carbon hydride metabolism, and two for amino acid metabolism. The result of our method well reflected the major role of liver, intermediate metabolism of amino acids.

2.2. *Cellular Component*

Table 4 shows top 20 TSF rank at liver on cellular component[c]. The last rank of TSF was 520th and the last rank of MEG was 9307th. In the table, three GO-terms appeared in both of top 20 TSF rank and top 20 MEG rank. They were 1) endocytotic vesicle (3rd TSF, 1st MEG), 2) peroxisome (6th TSF, 10th MEG) and 3) microsome (9th TSF, 5th MEG). The GO-terms in the both rank indicate that they were tissue specific and major transcripts in liver. In the followings, we validate

[b]The precise result is shown in Supplementary Table S4.
[c]the result of the top 20 MEG rank are available via our web site

Table 3. Number of Enzymes in Top20 TSF ranking and MEG ranking

tissue	num. enzymes TSF	MEG	tissue	num. enzymes TSF	MEG
adipose	6	3	lung	4	1
amnion	4	6	macrophage	2	5
brain	4	3	mammary gland	2	1
cerebellum	4	4	medulla oblongata	1	3
cerebral cortex	3	3	muscle	6	3
diencephalon	1	2	placenta	3	2
embryo	7	3	prostate gland	14	2
eye	3	2	somatosensory cortex	3	2
heart	7	4	striatal primordia	3	2
hippocampus	5	2	testis	6	3
liver	14	4	visual cortex	3	4

them. 1) Endocytotic vesicles are membrane-bound intracellular vesicles formed by invagination of the plasma membrane around an extracellular substance. As peroxisome and microsome are types of vesicles, we describe the role of peroxisome and microsome in the followings. 2) Peroxisome is a small, membrane-bound organelle that uses dioxygen to oxidize organic molecules, and contains enzymes that produce and others that degrade hydrogen peroxide. They fall under the term microbody, which holds the 6th TSF rank in Table 4, and were named after peroxidase-rich vesicles. Peroxidase held 3rd MEG rank in molecular function of liver (Table 2), indicating that peroxidase is highly produced in the liver. Table S6 shows the TSF ranks and number of normalized tags on peroxisome and peroxidase. From the table, peroxisome held a high TSF rank only in liver and also had the largest number of normalized tags in liver. The number of normalized tags in liver was three times greater than in heart, which had the second largest number of normalized tags. On the other hand, peroxidase in liver did not have particularly high TSF rank or number of normalized tags. Our result helped to find the fact that the presence of peroxisome is characteristic of liver, but peroxidase is not. 3) Microsomes are small vesicular particles containing high-density lipid. As mentioned above, the metabolism of lipid is a major role of the liver, and the mobilization and biosynthesis of triacylglycerol, a kind of lipid, were highly and locally occurred in liver. It indicates the GO-term microsome in liver were adequate to highly ranked.

The last two GO-terms we focused on were intercellular canaliculus (11th TSF) and connexon complex (18th TSF). These terms belongs to a GO-term named intercellular junction. Figure 1 shows a summary of intercellular junctions and related

Table 4. Top20 TSFs of liver on Cellular Component

rank TSF	rank MEG	Gene Ontology	#tag	#tag (tpm)
1	957	ornithine carbamoyltransferase complex	269	80.1
2	372	membrane attack complex	1295	385.4
3	1	endocytic vesicle	57969	17250.8
4	-	recombination nodule	4	1.2
4	7824	late recombination nodule	4	1.2
6	10	peroxisome	38807	11548.5
6	-	microbody	38807	11548.5
8	3110	glycine dehydrogenase complex (decarboxylating)	56	16.7
9	5	microsome	47059	14004.2
10	2219	vesicular fraction	47159	14033.9
11	669	intercellular canaliculus	1166	347.0
12	295	mitochondrial outer membrane translocase complex	797	237.2
13	8219	lateral element	3	0.9
14	922	citrate lyase complex	284	84.5
15	1926	glycine cleavage complex	174	51.8
16	1045	alpha-ketoglutarate dehydrogenase complex (sensu Eukaryota)	244	72.6
16	-	alpha-ketoglutarate dehydrogenase complex	244	72.6
18	362	connexon complex	1588	472.6
19	389	electron transfer flavoprotein complex (sensu Eukaryota)	1278	380.3
19	-	electron transfer flavoprotein complex	1278	380.3

GO-terms. The figure also shows the TSF rank of each GO-term in liver. The GO-term *Intercellular junction* has eight children terms, six of the eight terms had tags from liver, and two of the six and one descendant term held high TSF ranks. The three were *intercellular canaliculus* (11th TSF), *connexon complex* (19th TSF) and *gap junction* (27th TSF). On the other hand, the ranks of desmosome and tight junction were relatively low. A desmosome is a type of intercellular junction peculiar to epithelial cells, and a tight junction is found in epithelial cells and endothelial cells. The results match very well not only to the features of liver, but also to the feature of lung, heart, muscle and prostate gland.

Table 5 and Table S7 show the TSF ranks and number of normalized tags of the GO-terms in Figure 1 respectively. Lung had the highest TSF rank and largest

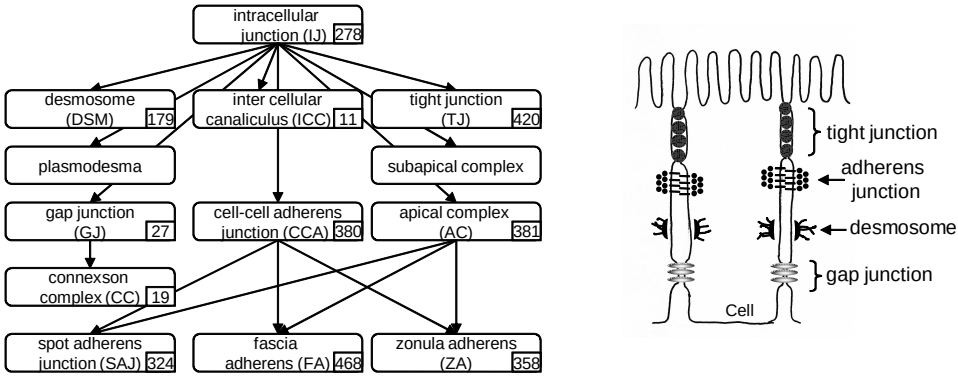

Fig. 1. Intercellular Junction and related GO terms and thier TSF ranks in liver. The arrows represent the relation of parent and child between GO terms. The number at lower right corner of each GO term is the TSF ranks in liver. GO terms without the number has no tags. Abbreviation of the terms are used in table 4 and S10.

number of normalized tags for desmosome and tight junction. What interested us was the tradeoff relationships between liver and lung in TSF rank. If the TSF rank of liver was higher than 30th TSF, the TSF rank of lung became lower, and vice versa except for fascia adherens. For fascia adherens, both liver and lung had low TSF ranks, 468th and 403rd TSF rank respectively. On the other hand, muscle, heart and prostate gland had very high TSF ranks, 1st, 4th and 10th TSF rank respectively. A fascia adherens is a broad intercellular junction in the intercalated disk of cardiac muscle that anchors actin filaments. As the cardiac muscle is a component of heart, the high rank in heart seemed appropriate. The tags of fascia adherens in heart, muscle and prostate gland were derived from one gene, the nebulin-related anchoring protein. (MGI:1098765). It had six GO-terms, actin cytoskeleton organization and biogenesis, fascia adherens, myofibril, actin binding, metal ion binding and protein binding. Although prostate gland has no cardiac muscle, smooth muscle is developed at connective cells around the glandular cells.

3. Discussion

CAGE is one of the newest technologies to measure the expression of genes. The largest advantage of CAGE is that it enables us to compare the expression of the gene in question to that of other genes in the same tissue. To leverage the advantage, we put forward a scheme to sum up the gene expression value along the hierarchical structure of gene ontology, and proposed a method to reveal the tissue specific functions by calculating the differential information content. The results shown in the former section were well-adapted to the feature of liver, especially in the case of intercellular junctions. Generally speaking, many genes shares one GO-term, especially in the category of cellular component, in their annotations. This circumstance prohibits usual methods to exhibit tissue specific cellular component because they

Table 5. TSF ranks and normalized tags of intercellular junction and related GO terms.

	TSF rank										
tissue	IJ	CCA	GJ	TJ	DSM	AC	ICC	SAJ	ZA	FA	CC
adipose	443	434	439	413	465	420	472	-	394	128	412
brain	361	419	454	105	418	218	449	-	457	-	463
cerebellum	132	442	26	38	330	83	423	41	503	358	255
cerebral cortex	221	337	281	68	-	111	295	-	-	-	322
diencephalon	227	371	343	125	258	195	378	-	-	-	358
embryo	362	279	401	488	82	443	479	225	485	197	372
heart	204	167	201	291	35	158	237	-	287	4	182
hippocampus	143	200	-	38	-	90	-	-	-	-	-
liver	278	380	27	420	179	381	11	324	358	468	18
lung	23	21	426	13	29	14	227	25	8	403	378
macrophage	361	190	474	469	486	443	450	-	317	490	475
mammary gland	35	20	-	-	-	-	-	-	-	-	-
medulla oblongata	116	206	73	53	-	111	-	-	-	-	-
muscle	284	164	368	418	123	384	373	-	416	1	382
prostate gland	189	364	318	72	206	116	420	1	378	10	270
somatosensory cortex	396	460	162	225	465	365	410	-	496	358	169
striatal primordia	262	190	321	327	244	240	333	-	90	-	301
testis	210	151	101	409	342	357	447	-	195	-	68
visual cortex	390	464	237	244	442	355	442	-	484	-	242

Note: Abbreviations use in the table is same as figure. 1

do not take countability of expression in to account. There is a point to be improved in our method. In the results, some GO-terms held same rank and they were often parent-child in the GO hierarchy. They could be redundant and may disturb the further analysis.

4. Conclusion

We presented a new method to describe tissue-specific function that leverages the advantage of CAGE. The method used molecular function of Gene Ontology terms to describe the tissue-specific functions and measure the tissue-specificities by the Information Content of the terms. As a by-product of using Gene Ontology, it gave us the information about tissue specificity on not only molecular functions, but also biological process and cellular component. The method was applied to the

CAGE data on 22 tissues, and TSF ranks and also MEG ranks were calculated. The majority of the results agree well with the well-known tissue roles. The results of our method will provide researchers a clue to the further research of the tissue roles and the deeper functions of the tissue-specific genes.

5. Materials and Methods

5.1. *Materials*

Mouse CAGE tag data, which includes 11,567,973 tags, were used as the gene expression profiles. The tags were derived from 23 types of tissues. As the 23rd tissue is UNDEFINED TISSUE TYPE, we eliminated it from the analyses. Tags with the same sequence were clustered to a representative tag, and representative tags were mapped to mouse mm5 genes or, namely, FANTOM3's Transcriptional Unit (TU) [17]. The TUs were annotated with controlled nomenclature vocabulary transferred from the original literature and/or GO terms by the curators. We eliminated genes that have only one tag in total among all tissue. They are regarded as in the range of errors [22]. All data is available at the FANTOM3 Web page, and the details are described in the FANTOM3 papers [17].

5.2. *Methods*

The tissue specificities of GO-terms are measured using the differential information content (DIC). The GO-terms are ranked by DIC, and are called TSF rank. Let $Exp[tissue][gene]$ be the number of tags of *gene* at *tissue*, $GO[gene]$ be a set of GO-terms that *gene* has as its annotation. The TSF rank is calculated as follows.

Step 1. Calculate the expression amount of GO-terms *term* for each *tissue*:

$$GO\_Exp[term][tissue] = \sum_{term \in \text{Ancestors of } GO[gene]} Exp[tissue][gene].$$

Step 2. Sum up the expression amount of all tissues for each GO-term *term*:

$$GO\_Exp[term][all] = \sum_{\text{all tissues}}^{tissue} GO\_Exp[term][tissue].$$

Step 3. Calculate DIC of each *tissue* for each GO-term *term*:

$$DIC[term][tissue] = -\log \frac{GO\_EXP[term][tissue]}{GO\_EXP[root][tissue]} - \left(-log \frac{GO\_EXP[term][all]}{GO\_EXP[root][all]} \right)$$

,where *root* is the root of GO, gene ontology.

Step 4. Rank GO-terms for each tissue according to descending order of DIC.

In the paper, we showd MEG ranks of GO-terms for each tissue. The MEG is ranked according to descending order of *MEG_SCORE*. It is calculated as follows:

$$MEG\_Score[term][tissue] = \sum_{term \in GO[gene]} Exp[tissue][gene].$$

References

[1] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, *Proc. Natl. Acad. Sci. USA* **9**, 14863 (1998).

[2] t. Akutsu, S. Kuhara, O. Maruyama and S. Miyano, *Genome Informatics* **9**, 151 (1998).

[3] N. Friedman, M. Linial, I. Nachman and D. Pe'er, *Journal of Computational Biology* **7**, 601 (2000).

[4] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Andres *et al.*, *Mol. Biol. Cell* **9**, 3273 (1998).

[5] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos *et al.*, *Science* **252**, 1651 (1991).

[6] V. E. Velculescu, L. Zhang, B. Volgelsteine and K. W. Kinzler, *Science* **270**, 484 (1995).

[7] S. Saha, A. B. Sparks, C. Rago, V. Akmaev, C. J. Wang *et al.*, *Nat. Biotechnol.* **20**, 508 (2002).

[8] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd *et al.*, *Nat. Biotechnol.* **18**, 630 (2000).

[9] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa *et al.*, *Proc. Natl. Acad. Sci. USA* **100**, 15776 (2003).

[10] R. Kodzius, Y. Matsumura, T. Kasukawa, K. Shimokawa, S. Fukuda *et al.*, *FEBS Lett.* **559**, 22 (2004).

[11] J. Schug, W. P. Schuller, C. Kappen, M. Salbaum, J. M. Bucan and C. J. Stoeckert Jr, *Genome Biology* **6:R33** (2005).

[12] K. Kadota, S. Nishimura, H. Bono, S. Nakamura, Y. Hayashizaki, Y. Okazaki and K. Takahashi, *Physiol Genomics* **12**, 251 (2003).

[13] K. Kadota, J. Ye, Y. Nakai, T. Terada and K. Shimizu, *BMC Bioinformatics* **7:294** (2006).

[14] Gene Ontology Consortium, *Genome Res.* **11**, 1425 (2001).

[15] Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi *et al.*, *Nature* **420**, 563 (2002).

[16] T. Imanishi, T. Itoh, Y. Suzuki *et al.*, *PLoS Biology* **2**, 856 (2003).

[17] P. Carninci, T. Kasukawa, S. Katayama *et al.*, *Science* **309**, 1559 (2005).

[18] P. Carninci, C. Kvam, A. Kitamura, T. Ohsumi, Y. Okazaki *et al.*, *Genomics* **37**, 327 (1996).

[19] P. Carninci and Y. Hayashizaki, *Methods Enzymol* **303**, 19 (1999).

[20] K. Ito (ed.), *Encyclopedic Dictionary of Mathematics* (The MIT Press, 1993), ch. 213. Information Theory, second edn.

[21] M. Kanehisa, *Science and Technology Japan* , 33 (1996).

[22] M. C. Firth, L. G. Wilming, A. Forrest, H. Kawaji *et al.*, *PLoS Genetics* **2** (2006).