Extraction of Organism Groups from Phylogenetic Profiles Using Independent Component Analysis

Yoshihiro Yamanishi[1] yoshi@kuicr.kyoto-u.ac.jp
Masumi Itoh[1] itoh@kuicr.kyoto-u.ac.jp
Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Abstract

In recent years, the analysis of orthologous genes based on phylogenetic profiles has received popularity in bioinfomatics. We propose a new method to extract organism groups and their hierarchy from phylogenetic profiles using the independent component analysis (ICA). The method involves first finding independent axes in the projected space from the multivariate data matrix representing phylogenetic profiles for a number of orthologous genes. Then the extracted axes are correlated with major organism groups, according to the extent of affiliaion of axes scores for all the genes to specific organisms. The ICA was applied to the phylogenetic profiles created for 2875 orthologs in 77 organisms by using the KEGG/GENES database. The 9 extracted components out of 18 predefined components well represented the organism groups as categorized in KEGG. Furthermore, we performed the cluster analysis and obtained the hierarchy of organism groups.

Keywords: phylogenetic profile, orthologous gene, independent component analysis, feature extraction

1 Introduction

A growing number of fully sequenced genomes makes it possible for us to conduct a large scale comparative genomic research. The availability of such data has accelerated the development of bioinformatics technologies to uncover the machinery of the cell. One of the important problems in bioinformatics is to predict biological functions of genes identified in the genome. Recently, a functional prediction method based on phylogenetic profiles has been proposed [**?**, **?**], where a phylogenetic profile is defined as a bit pattern that encodes the presence or absence of conserved (orthologous) genes in a set of organisms with fully sequenced genomes. When two genes share similar phylogenetic profiles, showing correlated patterns of presence or absence among the organisms, it is assumed that these genes are also functionally correlated. The Hamming distance, which is defined as the number of differences in the bit pattern, has been uses as a measure for evaluating the similarity between two profiles. Another interest is the construction of genome trees from the phylogenetic profiles. This idea stems from the assumption that gene losses or acquisitions are major evolution phenomena [**?**, **?**, **?**].

In this paper, we propose a new method based on the independent component analysis (ICA) to analyze phylogenetic profiles and to capture essential biological features. A set of phylogenetic profiles can be regarded as a multivariate data matrix, so ordinary methods in multivariate data analysis can be applied. ICA, which is one of the multivariate data analysis, has received attention in bioinformatics in recent years. For example, ICA has been used to analyze gene expression data, where Liebermeister *et al.* [**?**] refer the extracted independent components as "expression modes" corresponding to distinct biological functions. The purpose of the present study is to find independent components that characterize major organism groups and, at the same, to identify genes that are

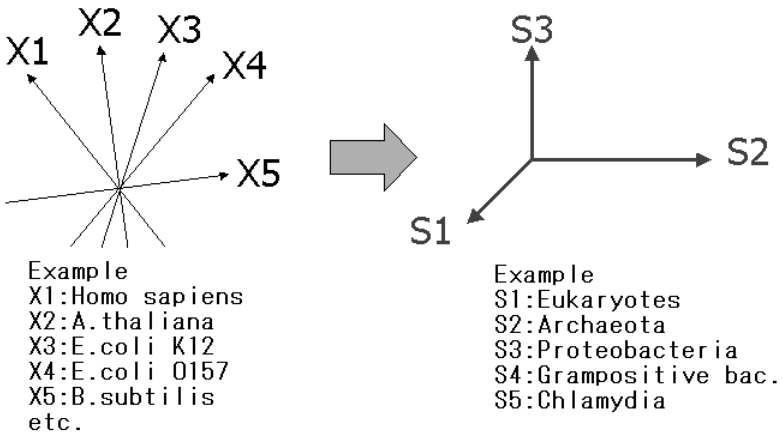

Figure 1: Projection from the organism space to a lower-dimensional space representing organism groups.

characteristic to each organism group. Here we apply ICA to the phylogenetic profiles created by using the KEGG database, develop a method to correlate independent components to organism groups, and investigate how ICA can be used in functional predictions.

2 Materials and Methods

2.1 Data

The phylogenetic profiles were constructed from 2875 orthologous genes in 77 organisms, as defined by the KEGG/GENES database [**?**, **?**] as of May 2002. In this study, we focus on the organisms with fully sequenced genomes, including 6 eukaryotes, 13 archaea, and 58 bacteria. Each phylogenetic profile consists of a string of bits, in which the presence and absence of an orthologous gene are coded as 1 and 0, respectively, across the above organisms.

2.2 Independent Component Analysis

It is important to find a suitable projection of multivariate data in many application. The independent component analysis (ICA) is a recently developed, linear transformation method in the field of statistics and signal processing [**?**, **?**, **?**]. The purpose of the method is to represent a set of variables as a linear combination of latent variables which are statistically independent each other. In actual applications, ICA has been used for blind source separation, feature extraction, redundancy reduction, blind deconvolution, and many more. Figure **??** shows an illustration of the current study, which involves projection from the high-dimensional organism space to a low-dimensional space whose axes are represented by the independent components.

Given a vector of observed variables $\mathbf{x} = (x_1, x_2, \cdots, x_P)^T$, where P is the number of variables, a mathematical model of ICA is formulated as

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where $\mathbf{s} = (s_1, s_2, \cdots, s_M)^T$ is the vector of statistically independent latent variables called independent components, M is the number of independent components, $\mathbf{A}$ is the unknown constant matrix. The vector $\mathbf{s}$ is called the independent component score (IC score), and calculated as $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x} = \mathbf{W}\mathbf{x}$,

so the estimation of $\mathbf{s}$ is equivalent to finding a weight matrix $\mathbf{W}$. The variance of each independent component cannot be defined, so we assume $E\{s_j^2\} = 1$.

The related methods to ICA are the principal component analysis (PCA) and the factor analysis (FA), which are well-known linear transformation methods. However, the concept of PCA and FA are different from that of ICA, although both can be used as methods to capture essential structures of multivariate data. PCA and FA are based on the assumption that latent variables such as eigenvectors or factors are uncorrelated and gaussian. On the other hand, ICA is based on the assumption that latent variables are not only uncorrelated but also statistically independent and non-gaussian. In general, the requirement of independence is mathematically much stronger than that of uncorrelatedness. Indeed, ICA behaves differently from PCA or FA in actual applications. The objective of finding such a set of statistically independent components is accomplished by maximizing the non-gaussianity of latent variables.

Suppose that we seek for one of the independent components as $y = \mathbf{w}^T\mathbf{x}$, and the negentropy, which is one of the popular measures of non-gaussianity, is used as a measure of non-gaussianity. The entropy of a discrete variable Y is defined as

$$H(Y) = -\sum_i P(Y = a_i) \log P(Y = a_i), \tag{2}$$

where a_i is a possible value of Y. In general, the differential entropy H of $\mathbf{y}$ with probability density function $f(\mathbf{y})$ is defined as

$$H(\mathbf{y}) = -\int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y}. \tag{3}$$

Then, the negentropy of $\mathbf{y}$ is defined as

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y}), \tag{4}$$

where $\mathbf{y}_{gauss}$ is a gaussian random vector of the same covariance matrix as $\mathbf{y}$. The estimation of independent components is reduced to finding $\mathbf{y}$ in such a way that maximizes $J(\mathbf{y})$.

2.3 Algorithm of ICA

Note that the negentropy is used as a measure of non-gaussianity of latent variables. In this case, the estimation of the independent components leads to finding directions in which the negentropy is maximized. The algorithm we adopt is FastICA (see Hyvärinen, 1999), in which non-gaussianity is measured by approximating the negentropy J. The approximation takes the form

$$J(y) \approx [E\{G(y)\} - E\{G(\nu)\}]^2. \tag{5}$$

where ν is the normalized gaussian variable, y is assumed to have zero mean and a unit variance, and $G(\cdot)$ is a non-quadratic function. Practical choices of $G(\cdot)$ have been proposed as follows:

$$G_1(u) = \log\cosh \alpha_1 u, \quad G_2(u) = \exp(-\alpha_2 u^2/2). \tag{6}$$

where α_1, α_2 are some suitable positive constants.

The FastICA algorithm for estimating $\mathbf{w}$ is formulated by the following Newton-like iteration using the derivative $g(\cdot)$ of $G(\cdot)$. A random vector $\mathbf{w}_0$ is chosen as an initial weight vector $\mathbf{w}$ in advance.

Step 1. $\mathbf{w} \leftarrow E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$.

Step 2. $\mathbf{w} \leftarrow \mathbf{w}/\parallel \mathbf{w} \parallel$.

Step 3. Go to Step 1 until a convergence criterion is met.

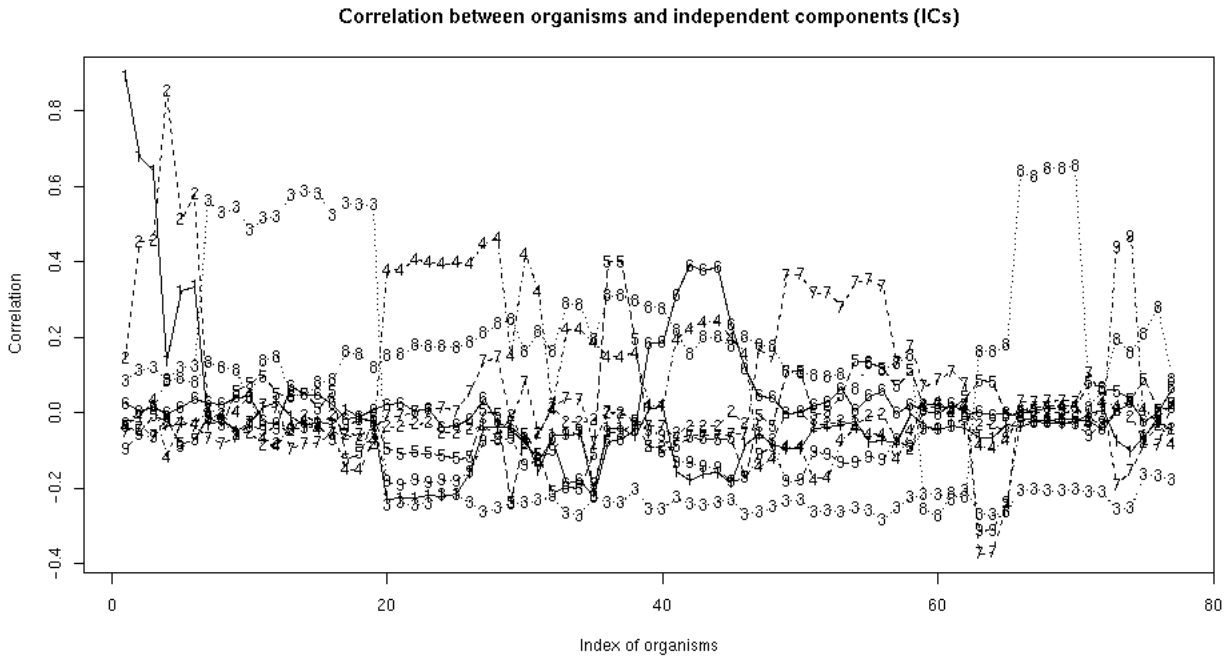

Figure 2: Correlation coefficients between organisms and ICs.

Here $\|\cdot\|$ indicates the Euclidean norm. This procedure is repeated for a given number of components applying the process of decorrelating the components each other. One technique for preventing a few weight vectors from getting privileges over others is the following symmetric decorrelation:

$$\mathbf{W} \quad \leftarrow \quad (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W}, \tag{7}$$

where $\mathbf{W}$ is the matrix $(\mathbf{w}_1, \cdots, \mathbf{w}_M)^T$ of the weight vectors.

2.4 Correlation of Independent Components and Organism Groups

Suppose that the variable X corresponds to an organism in the collection of phylogenetic profiles and the variable Y corresponds to an independet component. In order to evaluate if any of the independent components are related to specific groups of organisms, the following correlation coefficient is used:

$$r(X,Y) = \frac{1/N\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{(1/N\sum_{i=1}^{N}X_i-\bar{X})^2}\sqrt{(1/N\sum_{i=1}^{N}Y_i-\bar{Y})^2}}, \tag{8}$$

where N is the number of genes, $\bar{X}$ and $\bar{Y}$ are the means of the variables X and Y, respectively.

3 Results

3.1 Extraction of Independent Components

The collection of phylogenetic profiles was considered as a multivariate data matrix whose rows corresponded to individuals (genes) and whose columns corresponded to variables (organisms). We applied

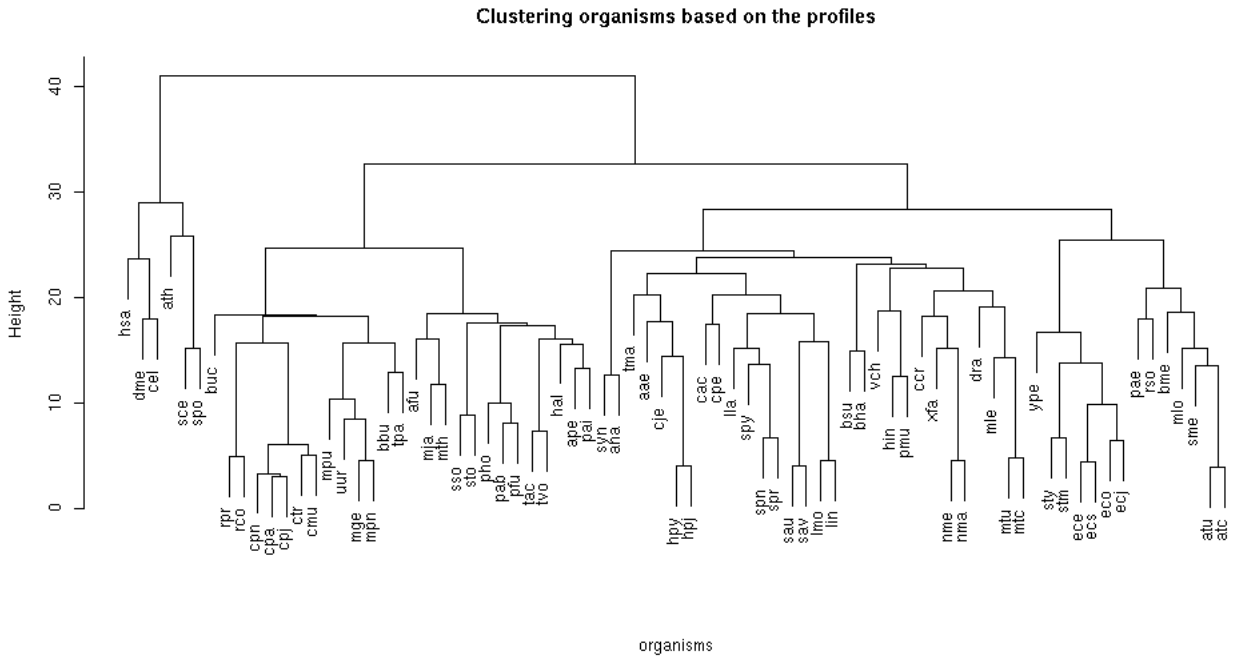

Figure 3: Clustering organisms based on original phylogenetic profiles

ICA to this data matrix by specifying the number of independent components to be extracted. We report here the result of using 18 independent components for the analysis of the data matrix with 2875 rows (genes) and 77 columns (organisms). To avoid an extra computational burden, we conducted a prewhitening process using eigenvalue decomposition. In most cases, this process is equivalent to computing the normalized principal components of the original data matrix. It is well-known that the prewhitening process in ICA has effects of reducing the dimension and removing the noise. In this study, we selected G_1 in eq.(6) as a non-quadratic function with $\alpha_1 = 1$, and we conducted the prewhitening process with 18 eigenvectors.

3.2 Grouping of Organisms

The original data matrix of 2875 rows (genes) and 77 columns (organisms) was converted by ICA to a data matrix of 2875 rows and 18 columns (independent components). To interpret biological meanings of the extracted independent components (ICs), we examined if any ICs could be correlated with groups of organisms by computing correlation coefficients for all combinations of 77 organisms and 18 ICs. We found 9 out 18 components were well correlated with specific organism groups. Figure **??** shows these correlations, where the vertical axis represents the correlation coefficient and the horizontal axis represents the organisms which are in the order of eukaryotes, archaea, and bacteria. The numbers 1 to 9 indicate the IC numbers, which were numbered manually along the order of the organisms. From this figure, we can see peaks and valleys of correlation coefficients, which indicates the existence of specific ICs with major contributions to specific organism groups. We compared a set of organisms corresponding to a peak or a valley with the organism groups (taxonomy) in KEGG. Table **??** summarizes the result of assigning each of the 9 ICs to a specific organism group. Among the 77 organisms used for the construction of phylogenetic profiles, 74 were well represented by the 9 ICs. The outliers were *Deinococcus raiodurans*, *Aquifex aeolicus*, and *Thermotoga maritima*.

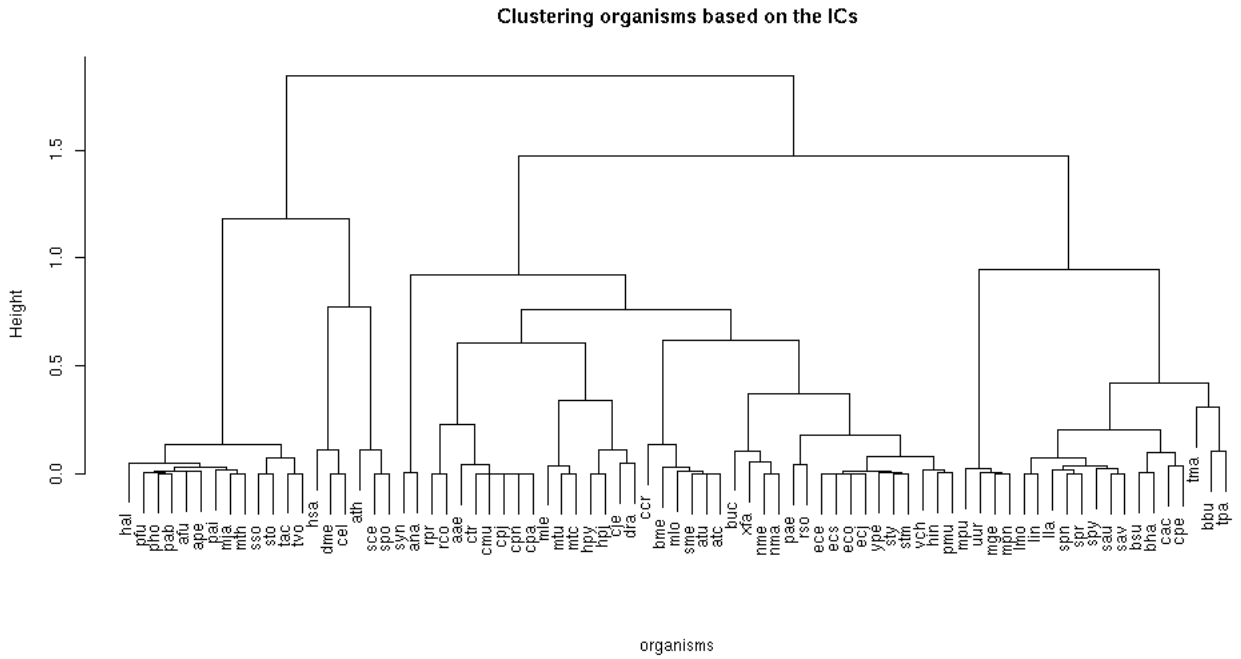

Figure 4: Clustering organisms based on the correlation coefficients between organisms and ICs

3.3 Hierarchy of Organism Groups

The multivariate data matrix representing the original data set of phylogenetic profiles can also be used to classify organisms. We defined the similarity by the Hamming distance and performed the complete linkage hierarchical cluster analysis. Figure **??** shows the result, where the labels for the leaves indicate the organism abbreviations in KEGG. In comparison, we also performed the hierarchical cluster analysis using the result of the IC analysis. We considered the 9-variable vector containing the values of correlation coefficients between IC scores and organisms shown in Figure **??** and defined the similarity between two vectors by, again, the correlation coefficient. When the result of Figure **??** is compared with Figure **??**, the hierarchy appeared much simpler and well corresponded to the simplified classification of KEGG organism groups, which is actually based on the NCBI taxonomy. When all the 18 ICs were used for the cluster analysis, the result was more complicated and became similar to Figure **??** (data not shown).

3.4 Genes Identified

The result of ICA can be used to identify genes that are clustered at high and low scores along each independent component. For the illustration of how the IC scores can be used, Figure **??** to **??** show the scatter plot of IC scores, IC2 vs. IC1, IC4 vs. IC3, IC6 vs. IC5, and IC8 vs. IC7, respectively. The IC scores in the projected space visualized the distribution of genes in the axes of assigned organism groups, and enabled us to highlight and detect some characteristic genes to certain organism groups. Table **??** shows the number of genes in every organism whose scores are higher than the 0.05 percentile in each IC. It is clear, for example, that the IC1 axis that characterizes the animal group is dominated by a number of animal genes, but at the same time there are also non-animal genes. In fact, some non-animal genes have higher IC scores than animal genes, because they are exclusively found in non-animal groups.

Table 1: Interpretation of extracted independent components (ICs).

	Organism group	Examples
IC1	Eukaryotes(animal)	Homo sapiens, Drosophila melanogaster, etc.
IC2	Eukaryotes(plant/fungi)	Arabidopsis thaliana, budding yeast, etc.
IC3	Archaea	Methanococcus jannaschii, Thermoplasma acidophilum, etc.
IC4	Proteobacteria(gamma)	Escherichia coli, Salmonella typhi, etc.
IC5	Proteobacteria(delta/epsilon)	Helicobacter phlori, Ralstonia solanacearum, etc.
IC6	Proteobacteria(alpha)	Mesorhizobium loti, Sinorhizobium meliloti, etc.
IC7	Grampositive bacteria (Low G+C)	Bucillus subtilis, Bucillus halodurans, etc.
IC8	Chlamydia	Chlamydia trachomaticm, Chlamydia muridarum, etc.
IC9	Cyanobacteria	Synechocys sp., Anabaena sp.

To confirm the validity of the extracted ICs and their assignments to organism groups from the viewpoint of biological functions, we examined the positions of selected genes in the KEGG biological pathways. For instance, we focus on the 4-th IC scores of genes. Recall that the IC4 corresponds to the Proteobacteria gamma group. Some high scoring genes in the IC4 are mapped to the lipopolysaccharide biosynthesis pathway. The genes marked with bold lines in Figure **??** show the high scoring genes in the IC4. These genes are in fact found in *Escherichia coli*, which of course belongs to the Proteobacteria gamma group, as shown by gray color in Figure **??**.

4 Discussion

In this study, we proposed to use the independent component analysis (ICA) for extraction of organism groups from phylogenetic profiles. We showed that extracted independent components (ICs) well corresponded to major organism groups. In the space projected by the ICA, we could detect genes which were characteristic (e.g., exclusively present or exclusively absent) to the associated organism groups.

Another possible approach was to use the principal component analysis (PCA). We applied PCA to the same phylogenetic profiles, but we could not extract biologically meaningful features except that the first principal component, which indicated the separation of prokaryotes and eukaryotes, and a few others. Compared to ICA, each PCA component captures a maximal variance of data assuming gaussianity. It is claimed that this constraint of maximizing variance tends to interrupt the process of detecting biologically meaningful features [**?**]. In contrast, ICA attempts to maximize nongaussianity, which is defined in the present analysis by the negentropy or, roughly speaking, the information content. Thus, ICA is an appropriate method to detect biological features.

In our problem, there exsists complex phylogenentic association, or the phylognetic tree structure, among the organism groups to be extracted. The similarity of two phylogenetic profiles is usually defined by the Hamming distance or the correlation coefficient. In both cases, however, the distance does not incorporate any tree structure. Vert has proposed a new similarity measure between two profiles by incorporating the phylogenetic tree among organisms and a mathematical framework using kernel methods [**?**]. His method improved the accuracy of allocating genes to a given number of functional groups. Although we attempted to identify hierarchy of organism groups, such information should be included in the process of extraction. The development from the "independent" components to "tree" components is our future work.

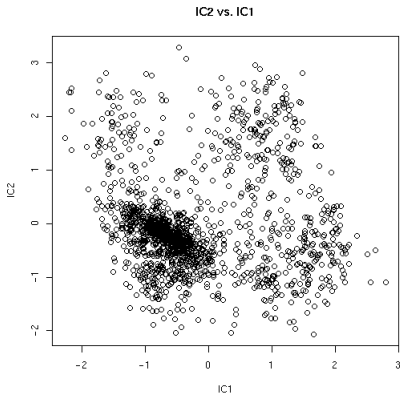

Figure 5: IC scores: IC2 vs. IC1

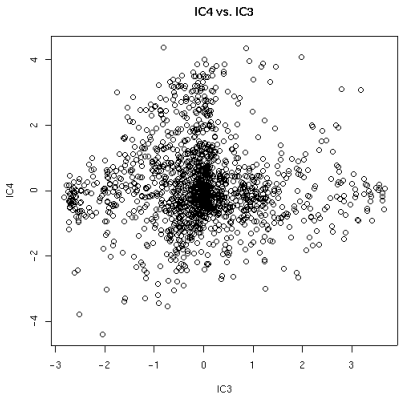

Figure 6: IC scores: IC4 vs. IC3

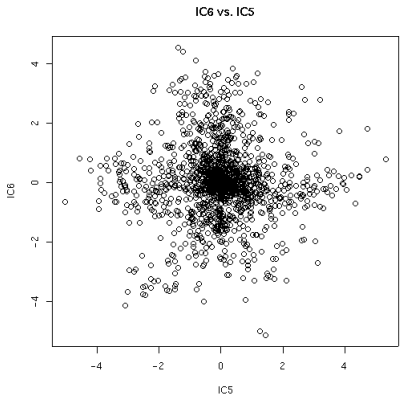

Figure 7: IC scores: IC6 vs. IC5

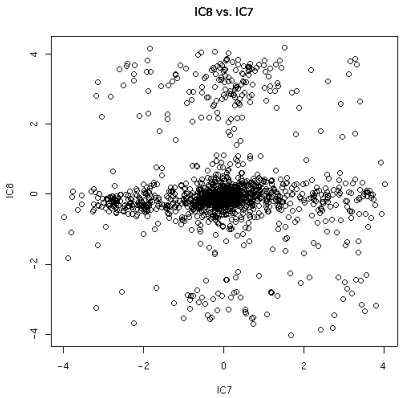

Figure 8: IC scores: IC8 vs. IC7

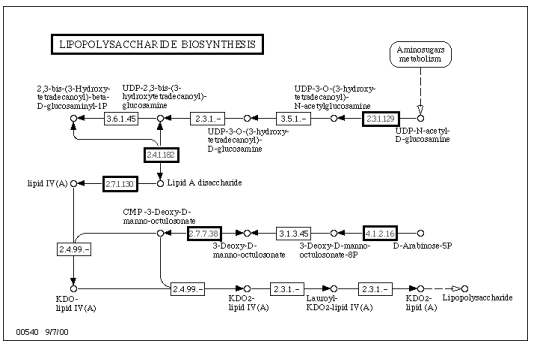

Figure 9: An example of pathways: the genes marked with bold lines have high scores in IC4.

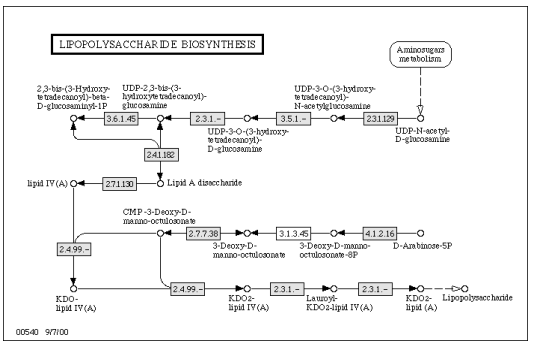

Figure 10: An example of pathways: the genes colored in gray are the ones E.coli have.

5 Acknowledgements

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Comon, P., Independent component anlaysis - a new concept?, *Signal Processing*, 36:287-314, 1994.

[2] Hyvärinen, A., Fast and robust fixed-point algorithms for independent component analysis, *IEEE trans. on Neural Networks*, 10(3):626-634, 1999.

[3] Hyvärinen, A., Survey on independent component analysis, *Neural Computing Surveys*, 2:94-128, 1999.

[4] Lin, J. and Gerstein, M., Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels, *Genome Res.*, 10:808-818, 2000.

[5] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A., The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30:42-46, 2002.

[6] Liebermeister, W., Linear modes of gene expression determined by independent component analysis, *Bioinformatics*, 18(1):51-60, 2002.

[7] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., Eisenberg, D., A combined algorithm for genome-wide prediction of protein function, *Nature*, 402, 83-86, 1999.

[8] Pellegrini, M., Marcotte, E.M., Thompson, M. J., Eisenberg, D., and Yeates, T.O. Assigning protein functions by comparative geneme analysis: Protein phylogenetic profiles, *Proc. Natl. Acad. Sci. USA*, 96:4285-4288, 1999.

[9] Tekaia, F., Lazcano, A., and Dujon, B., The genomic tree as revealed from whole proteome comparisons, *Genome Res.*, 9:550-557, 1999.

[10] Vert, J.-P., A tree kernel to analyze phylogenetic profiles, *Bioinformatics*, 18:S276-S284, 2002.

[11] Wolf, Y.I., Rogozin, I.B., Grishin, N.V. and Koonin, E.V., Genome trees and the tree of life, *Trends in Genetics*, 18:9:472-479, 2002.

[12] http://www.genome.ad.jp/kegg/

Table 2: The number of high scoring genes ($s_j > 0.05$ percentiles) in ICs for every organism.

No.	Abbr.	Organism	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9
1	hsa	Homo sapiens	143	23	64	28	29	41	33	33	18
2	dme	Drosophila melanogaster	143	30	63	23	23	27	23	31	18
3	cel	Caenorhabditis elegans	139	34	59	24	28	26	25	30	14
4	ath	Arabidopsis thaliana	1	48	34	9	20	18	18	40	23
5	sce	Saccharomyces cerevisiae	91	38	70	24	34	26	32	53	15
6	spo	Schizosaccharomyces pombe	84	42	66	20	28	21	26	46	16
7	mja	Methanococcus jannaschii	8	5	133	16	36	17	17	49	22
8	mth	Methanobacterium thermoautotrophicum	8	8	131	19	34	18	16	52	18
9	afu	Archaeoglobus fulgidus	11	6	136	23	54	29	26	59	26
10	hal	Methanopyrus kandleri	13	6	119	27	45	32	32	50	15
11	tac	Archaeoglobus fulgidus	15	1	117	6	42	13	27	49	15
12	tvo	Thermoplasma volcanium	14	2	117	6	35	13	26	47	15
13	pho	Pyrococcus horikoshii	7	2	129	8	41	22	18	30	14
14	pab	Pyrococcus abyssi	7	6	137	14	45	23	23	39	16
15	pfu	Pyrococcus furiosus	9	4	138	8	40	22	22	45	18
16	ape	Aeropyrum pernix	10	5	119	13	38	22	26	51	16
17	sso	Sulfolobus solfataricus	16	2	132	5	23	15	26	51	23
18	sto	Sulfolobus tokodaii	13	4	131	4	27	16	25	53	20
19	pai	Pyrobaculum aerophilum	14	0	130	11	37	19	32	56	22
20	eco	Escherichia coli K-12 MG1655	33	24	31	142	106	86	93	121	48
21	ecj	Escherichia coli K-12 W3110	33	23	31	142	96	86	91	121	47
22	ece	Escherichia coli O157 EDL933	33	26	31	141	102	85	94	130	49
23	ecs	Escherichia coli O157 Sakai	33	26	31	141	100	84	92	129	46
24	sty	Salmonella typhi	32	24	33	140	95	76	95	127	48
25	stm	Salmonella typhimurium	33	24	34	141	94	77	93	128	48
26	ype	Yersinia pestis	31	23	27	136	88	74	91	128	39
27	hin	Haemophilus influenzae	32	14	13	127	75	54	67	105	23
28	pmu	Pasteurella multocida	32	11	15	138	72	45	74	114	29
29	xfa	Xylella fastidiosa	22	11	14	62	29	30	39	98	25
30	vch	Vibrio cholerae	36	16	31	132	79	48	88	117	37
31	pae	Pseudomonas aeruginosa	34	19	26	120	87	53	71	131	43
32	buc	Buchnera sp. APS	13	8	10	18	11	6	26	56	7
33	nme	Neisseria meningitidis MC58 (serogroup B)	20	14	14	84	50	10	50	109	27
34	nma	Neisseria meningitidis Z2491 (serogroup A)	21	14	14	85	49	12	50	109	29
35	rso	Ralstonia solanacearum	28	24	28	83	82	23	64	121	34
36	hpy	Helicobacter pylori 26695	21	15	15	53	126	16	29	98	24
37	hpj	Helicobacter pylori J99	21	15	14	53	123	16	27	98	24
38	cje	Campylobacter jejuni	20	9	19	61	89	18	28	105	29
39	rpr	Rickettsia prowazekii	15	3	6	32	10	36	13	76	13
40	rco	Rickettsia conorii	18	2	6	32	11	38	13	76	14
41	mlo	Mesorhizobium loti	33	21	25	84	80	123	64	125	39
42	sme	Sinorhizobium meliloti	30	20	23	98	87	138	68	116	45
43	atu	Agrobacterium tumefaciens C58 (UWash/Dupont)	29	21	20	94	84	140	65	118	37
44	atc	Agrobacterium tumefaciens C58 (Cereon)	29	19	20	94	82	141	65	118	37
45	bme	Brucella melitensis	29	28	23	85	77	100	63	112	34
46	ccr	Caulobacter crescentus	25	13	15	68	56	60	51	105	22
47	bsu	Bacillus subtilis	34	18	19	17	82	55	102	108	24
48	bha	Bacillus halodurans	33	14	21	14	76	56	94	107	30
49	sau	Staphylococcus aureus N315 (MRSA)	26	17	16	14	83	31	137	89	7
50	sav	Staphylococcus aureus Mu50 (VRSA)	26	17	16	14	81	31	136	87	7
51	lmo	Listeria monocytogenes	28	13	17	8	67	30	115	82	17
52	lin	Listeria innocua	28	13	17	9	68	30	110	81	16
53	lla	Lactococcus lactis	27	12	12	12	61	34	91	72	12
54	spy	Streptococcus pyogenes SF370 (serotype M1)	24	10	15	15	59	19	100	62	6
55	spn	Streptococcus pyogenes MGAS8232 (serotype M18)	21	10	16	16	60	29	104	81	15
56	spr	Streptococcus pneumoniae R6	21	10	8	17	61	32	103	74	15
57	cac	Clostridium acetobutylicum	23	11	20	17	82	38	81	84	24
58	cpe	Clostridium perfringens	29	6	25	22	71	35	81	86	20
59	mge	Mycoplasma genitalium	8	4	4	9	15	9	25	0	4
60	mpn	Mycoplasma pneumoniae	8	4	3	11	19	9	31	0	4
61	mpu	Mycoplasma pulmonis	11	4	3	12	20	15	31	0	6
62	uur	Ureaplasma urealyticum	7	7	3	11	21	7	25	0	6
63	mtu	Mycobacterium tuberculosis H37Rv (lab strain)	28	17	22	23	85	30	5	93	0
64	mtc	Mycobacterium tuberculosis CDC1551	27	16	22	22	85	28	4	92	0
65	mle	Mycobacterium leprae	23	5	9	16	46	19	0	81	0
66	ctr	Chlamydia trachomatis	14	6	11	18	18	13	18	143	13
67	cmu	Chlamydia muridarum	15	6	11	18	17	12	17	141	13
68	cpn	Chlamydophila pneumoniae CWL029	16	6	11	18	23	14	20	143	14
69	cpa	Chlamydophila pneumoniae AR39	15	6	11	18	20	14	20	143	14
70	cpj	Chlamydophila pneumoniae J138	16	6	11	18	22	14	20	143	14
71	bbu	Borrelia burgdorferi	16	7	12	15	28	7	32	30	6
72	tpa	Treponema pallidum	17	9	13	23	28	13	17	39	9
73	syn	Synechocystis sp. PCC6803	30	24	21	46	75	32	35	102	142
74	ana	Anabaena sp. PCC7120 (Nostoc sp. PCC7120)	28	21	24	47	75	40	43	100	143
75	dra	Deinococcus radiodurans	28	11	26	31	81	34	43	96	24
76	aae	Aquifex aeolicus	24	10	27	30	55	22	23	101	30
77	tma	Thermotoga maritima	25	9	23	14	49	35	44	68	24