Identification of Master Regulator Candidates for Diabetes Progression in Goto-Kakizaki Rat by a Computational Procedure

Shigeru Saito*‖, Yidan Sun†‖, Zhi-Ping Liu‡‖, Yong Wang§,
Xiao Han†, Huarong Zhou‡**, Luonan Chen‡** and Katsuhisa Horimoto§**
*INFOCOM Corporation, Tokyo 150-0001, Japan
†Key Laboratory of Human Functional Genomics of Jiangsu Province, Nanjing Medical University, Nanjing 210029, China
‡Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for Pre-Diabetes,
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China
§Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
¶Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology,
Tokyo 135-0064, Japan
‖These authors equally contributed to this work
**Corresponding authors: hrzhou@sibs.ac.cn, lnchen@sibs.ac.cn, k.horimoto@aist.go.jp

***Abstract*—Recently, we have identified 39 candidates of active regulatory networks for the diabetes progression in Goto-Kakizaki (GK) rat by using the network screening, which were well consistent with the previous knowledge of regulatory relationship between transcription factors (TFs) and their regulated genes. In addition, we have developed a computational procedure for identifying transcriptional master regulators (MRs) related to special biological phenomena, such as diseases, in conjunction of the network screening and inference. Here, we apply our procedure to identify the MR candidates for diabetes progression in GK rat. First, active TF-gene relationships for three periods in GK rat were detected by the network screening and the network inference, in consideration of TFs with specificity and coverage, and finally only 5 TFs were identified as the candidates of MRs. The limited number of the candidates of MRs promises to perform experiments to verify them.**

***Index Terms*—Master regulator; regulatory network; diabetes progression; systems biology.**

I. Introduction

Recent developments in genome-wide computational analyses successfully identified causal interactions [1], and showed promise in the identification of dysregulated genes within developmental and tumour-related pathways [2]. For example, a computational analysis procedure was applied to identify the MRs causally linked to the activation of a specific gene set, mesenchymal gene expression signature (MGES), in human malignant glioma [3]. Indeed, 53 TFs were obtained by ARACNe algorithm and the MGES enrichment test, and among them, the top 6 TFs with the largest fraction of MGES genes were experimentally controlled, as the MR candidates. Finally, 2 of the top 6 TFs, STAT3 and CEBPB, were experimentally verified as MRs of mesenchymal transformation. Unfortunately, the computational part in the previous work seems not so sophisticated and remains to be improved. For example, it is unclear why they selected the top 6 TFs from 53 TFs, not 5 or 7 TFs. Although the coverage of TFs for the MGES genes were carefully considered, there was no rational criterion at the final selection of the MR candidates. Furthermore, ARACNe considers the relationship between the three genes for selecting MR candidates. Actually, some mathematical techniques that can consider multiple relationships are well known, and are applied to infer the regulatory networks [4].

Recently, we developed a procedure for identifying MRs, by a combination of network screening and inference. The performance of our procedure was tested for MRs in human malignant glioma, by using the same data set [3]. Fortunately, our procedure worked well [5]. 22 TFs and 27 TFs were detected by the network screening and the inference, respectively, and 3 TFs overlapped between them. Interestingly, 2 of 3 TFs were STAT3 and CEBPB that were verified by experiments as the master regulators in the previous report.

In our previous paper [6], we have reported 39 candidates of active networks for the diabetes progression in GK rat, which were identified by the network screening, in comparison with the Wistar-Kyoto (WKY) rat. The candidates were characterized by the known biological pathways that were well consistent with the previous knowledge about the diabetes. Unfortunately, it was still insufficient to verify the plausibility of the active networks by experiments. This is partly because the results were presented as a metaphysical form, the biological pathway, instead of the list of concrete target genes, and partly because the active networks were composed of many genes that were not feasible for the experimental verification.

Here, we identify the candidates of master regulators for the diabetes progression in GK rat. Based on the networks specific to diabetes progression in our previous results [6], we tried to further narrow down the candidate molecules responsible for the diabetes by identifying the master regulators that play a central role for the diabetes progression in GK

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

rat. Furthermore, we improved our previous method [5], to consider the coverage of TF for its regulated genes in a statistical way, in addition to the specificity of TF to the target biological phenomena. As expected from the previous case of computational identification of MRs in human brain tumor [5] and the present improvement, the limited number of MRs was identified to give a hint to design further experimental works for candidate verification.

II. Materials and Methods

A. Overview of our procedure

Here, we searched for MR candidates by two approaches, which are schematically shown in Fig. 1. One is a knowledge-based approach, which estimates the consistency of the network structures among the known networks with the measured data (named "network screening") [6], [7]. Unfortunately, our knowledge about the gene variety in transcriptional networks is restricted. To compensate for this restriction, we use another approach to search for MRs, based on the inference of the network structures by using the measured data (path consistency algorithm) [5]. In both cases, we further select the MR candidates by considering the enrichment of gene expression signature in the networks.

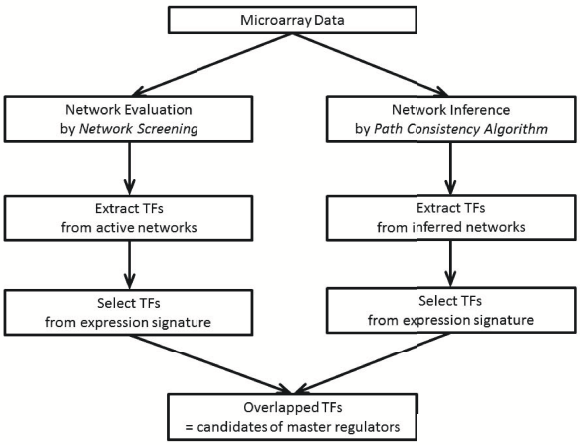

Fig. 1. Workflow of MR identification procedure.

B. Network screening

The candidates of active regulatory networks are detected by network screening in the following manner [6], [7]. First, the regulatory network sets are generated. The mouse binary relationships compiled in the TRANSFAC database [8] were used: based on the correspondence between mouse and rat in gene id, 3,015 binary relationships of 1,507 genes between 503 TFs and 1,123 regulated genes. Based on the binary relationships, transcriptional networks were constructed, according to the functional gene sets previously defined in the Molecular Signatures Database (MSigDB) [9]. In each gene set, the regulated genes in the binary relationships were searched, and if at least one gene was found in the gene set, then the corresponding binary relationships were regarded as a regulatory network characterized by the gene set. The set of constructed networks was used as the reference network for network screening. In present study, the reference network comprised 1,760 regulatory networks characterized by biological functions that are composed of 1,195 genes: the numbers of TFs and regulated genes are 335 and 860, respectively.

Then, we calculated the graph consistency probability (GCP) [10], which expresses the consistency of a given network structure with the monitored expression data of the constituent genes in this study. The details of the reference network and the GCP are described below.

First, suppose a causal graph is a directed acyclic graph (DAG), $G(V_i, E_j)$, where V_i is a vertex ($i = 1, 2, ..., n_v$) and E_j is an edge ($j = 1, 2, ..., n_e$) in the graph. The DAG can be factorized into subgraphs according to the parent-descent relationships. The joint density function $f(X_i)$, corresponding to V_i for the graph G, can then be factorized into the conditional density functions according to the graph, as follows:

$$f(X_1, X_2, ..., X_{n_v}) = \prod_{i=1}^{n} f(X_i|pa\{X_i\}), \qquad (1)$$

where $pa\{X_i\}$ is the set of variables corresponding to the parents of V_i in the graph.

Second, the causal graph meets the measured data based on the Gaussian graphical model (GN: Gaussian Network). On the assumption that the probability variable X_i is subjected to a multiple normal distribution, each conditional function in equation (1) is obtained by linear regression for the measured data of the constituent nodes (molecules) measured at m points, i.e.,

$$f(X_i|pa\{X_i\}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} exp\Big[-\frac{1}{2\sigma_i^2}\sum_{k=1}^{m}(x_{ik}-\sum_{j=1}^{n_i}\beta_{ij}x_{jk})^2\Big], \qquad (2)$$

where x_{ik} is the measured value of Xi, at the k-th point, and n_i is the number of variables corresponding to the parents of V_i. Thus, the joint density function in equation (1) is expressed by the regression for the measured data in equation (2). Finally, the logarithm of the likelihood of equation (2) is calculated for the measured data, as

$$\begin{aligned} l(G_0) &= ln\prod_{i=1}^{n_v} f(X_i|paX_i) \\ &= -\frac{1}{2}\sum_{i=1}^{n_v}\sum_{j=1}^{n_i}\{\frac{1}{\sigma_i^2}\sum_{k=1}^{m}(x_{ik}-\beta_{ij}x_{kj})^2 + ln(2\pi\sigma_i^2)\}. \end{aligned} \qquad (3)$$

Thus, the GN allows us to quantify a given network into the corresponding numerical value from the measured data, according to the network form. Note that the calculation of the likelihood itself requires no assumptions on the relationships between variables. Indeed, the likelihood can be calculated in the case of non-linear regressions, such as spline regression.

Finally, the probability of the log-likelihood for the network structure (graph consistency probability; GCP) was estimated

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

by the distribution of log-likelihoods for many networks, generated under the condition that the networks shared the same numbers of nodes and edges as those of the given network. Thus, we generated Nr networks under the same condition, and the GCP is simply defined, as

$$GCP = \frac{N_s}{N_r}, \tag{4}$$

where N_r is the total number of generated networks, and N_s is the number of networks with larger log-likelihoods than the log-likelihood of the tested network. In the present study, N_r was set to 2,000. In this paper, the GCP significance of the given network was set at 0.05 in this analysis.

C. Path consistency algorithm

The path consistency (PC) algorithm [11] is an algorithm to infer a causal graph composed of two parts: the undirected graph inference by a partial correlation coefficient and the following directed graph construction by the orientation rule. The present method partially exploits the first part of the PC algorithm for the inference of the network structures. A simple example in the PC algorithm is illustrated in Fig. 2.

We assume that five variable, X_1, X_2, X_3, X_4, X_5, have the five following relationships: i) $X_1 \coprod X_1$, ii) $X_2 \coprod (X_1, X_4)$, iii) $X_3 \coprod X_4 | (X_1, X_2)$, iv) $X_4 \coprod (X_2, X_3) | X_1$, and v) $X_5 \coprod (X_1, X_2) | (X_3, X_4)$. The PC algorithm reconstructs the above relationships as the follows. 1) Prepare a complete graph, C, between the five variables. 2) Test the correlation between two variables by calculating the zeroth-order of partial correlation coefficient (Pearson's correlation coefficient). From the test, two variables pairs, (X_1, X_2) and (X_2, X_4), are excluded (broken lines in Fig. 2), due to the relationships, i) and ii). 3) Test the correlation between three variables by calculating the first-order of partial correlation coefficient of variable pairs given one variable. Then, one variables pair, (X_3, X_4), is further excluded from the undated graph by 2), due to iii) and iv). 4) Test the correlation between four variables by calculating the second-order of partial correlation coefficient of variable pairs given two variables. Then two variables pairs, (X_1, X_5) and (X_2, X_5), are excluded, due to iv). 5) We cannot find any edges adjacent to three edges in the updated C. Thus, the algorithm naturally stops. As seen in the final graph, the five relationships emerged completely.

In general, the $(m-2)-th$ order of the partial correlation coefficient is calculated between two variables, given $(m-2)$ variables, i.e., $r_{ij,rest}$, between X_i and X_j, given the 'rest' of the variables, X_k for $k = 1, 2, ..., m$, and $k \neq i, j$, and after calculating the $(m-2)-th$ order of the partial correlation coefficient, the algorithm naturally stops. However, the algorithm does not usually request the $(m-2)-th$ order of the correlation coefficient for the natural stop. This is because adjacent variables, after excluding the variables, are often not found, even in the calculation of the lower orders of partial correlation coefficients.

In the actual expression profile data, many genes frequently show profiles with similar patterns. This makes the numerical calculation of correlation coefficients difficult, due to the multi-colinearity between the variables. The original PC algorithm accidentally stops, if only one correlation between a pair of variables shows a violation of the numerical calculation. However, in a biological sense, the gene pairs that cause the accidental stop can be interpreted as a case when they are highly associated with each other, in terms of gene expression. Thus, we modified the original PC algorithm to prevent it from accidentally stopping with the highly associated gene pairs, as follows [12], [13]. If the calculation of any order of the partial correlation coefficient between the variables is violated, then the corresponding pair of variables is regarded as being dependent. For example, if the first-order correlation coefficient, $r_{ij,k}$, cannot be calculated numerically, due to the multi-colinearity between X_i and X_j, then keep the edge $X_i - X_j$ without the statistical test. The other parts remain unchanged in the modified algorithm. Note that the above modification ensures that the algorithm will naturally stop for the data including the high correlation.

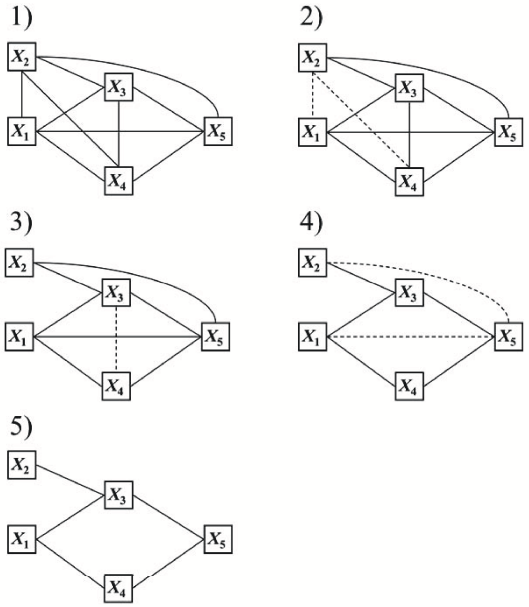

Fig. 2. Example of path consistency algorithm.

As seen in the original algorithm, the output is not unique, depending on the calculation order of pairs [11]. As a convenient way, a permutation test for the calculation order will be one of the ways to partly resolve this issue. In this study, the estimation without permutation was empirically adopted as a first approximation, based on the successful estimations of the relationships in our previous studies [12], [13]. In addition, one of the most remarkable features of PC algorithm is that the algorithm removes pseudo-correlations between the variables (genes) by considering the higher-order of partial correlations. If we have the measurement data for a complex network, we frequently face more serious issue on the pseudo-correlation rather than on the correlation level. The merit of PC algorithm

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

may be useful for identifying real relationships between TFs and their regulated genes.

D. Definition of MR by network screening and network inference

We first refer to two sets of networks obtained by the network screening [6], [7] and the network inference [12], [13]. From each set of network, the binary relationships between TFs and its regulated genes are extracted, only if the regulated genes are included in the expression signature, which is the ensemble of gene with the significant difference of gene expression that is statistically estimated by false discovery rate (FDR) test for multiple comparisons ($FDR < 0.05$) [14].

Then, we define MR candidates from the binary relationships by two criteria. One is the specificity of TF that is the same criterion as the previous method [5], and the other is the coverage of TF that is newly introduced in the present MR candidate identification. Here, the specificity simply means that TF emerged only in GK networks, but not in WKY. To select TFs in terms of the specificity, we select TFs that emerge at three periods in GK but not in WKY, as the MR candidates. The coverage means how many genes each TF regulates. To select TFs in terms of the coverage, we first counted the genes regulated by each TF for each period in GK and WKY, and then sort the numbers of regulated genes for each case. To consider the coverage in a rational way, we use the outlier test, Smirnov-Grubbs test [15], for the numbers of regulated genes, by setting a threshold ($p < 0.05$). Thus, TFs with the larger number of regulated genes that fulfill the threshold are selected in a statistical way. Finally, the two sets of MR candidates that are selected in terms of the specificity and the coverage are compared to define the final MR candidates.

1) Data analyzed in this study: We analyzed the gene expression data measured in GK and WKY rats [16], which is cited from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/projects/geo/) database (GSE:13271). The data are composed of 31,099 probes measured by using Affymetrix Microarray Suite 5.0 (Affymetrix), which are reduced into 14,506 genes, for 5 samples of male GK spontaneously diabetic rats and WKY rats at each of 5 time points (4, 8, 12, 16, and 20 weeks of age). In this analysis, the 5 periods are classified in to three periods: period of 4w, periods of 8w and 12w, and periods of 16w and 20w.

III. Results and Discussion

A. MR candidates detected by network screening

We identified 39 networks for GK and WKY rats in three periods of 4w, from 8w to 12w, and from 16w to 20w, by network screening, among the 1,760 networks in the reference network set in our previous study [6]. From the 39 networks, in total, we extracted 568 binary relationships of TF and its regulated gene, which were specifically found in the three periods for GK and WKY rats, under the condition that the gene expression shows difference with FDR of less than 0.05, between two kinds of rats for each period. The numbers of specific genes for each period in GK and WKY rats are as follows: 54 at the period of 4w in GK; 199 at 8w and 12w in GK; 56 at 16w and 20w in GK; 95 at the period of 4w in WKY; 125 at 8w and 12w in WKY; and 39 at 16w and 20w in WKY. Note that some TF-gene relationships emerged iteratively for different periods in GK and WKY rats.

First, the TF-gene relationships were selected by the specificity. To do this, we selected TFs that were listed in GK but not in WKY. Finally we found 21 TFs that regulates 32 genes in Table I. Here, all of the gene names are cited from the Rat Genome Database (http://rgd.mcw.edu/).

TABLE I
TFs identified by network screening in terms of specificity.

AR, BCL6, BRCA1, ETV4, FUS, GLI1, HES1, HNF1B, HNRNPK, KLF10, KLF4, LYL1, MEF2C, NFIA, NR2F1, NRL, PAX6, SP2, SP4, TCFAP2B, WT1

Next, the TF-gene relationships were selected by the coverage. We selected MR candidates by setting each threshold for each period in GK and WKY in Table II. As seen in the table, most TFs emerged in both GK and WKY, and finally we found 3 TFs (EGR1, NRF1, and TCFAP2A) that regulates 17 genes.

TABLE II
TFs identified by network screening in terms of coverage. ♯ means the number of regulated genes. TFs found in both GK and WKY are indicated by bold letters.

4w				8w_12w				16w_20w			
GK		WKY		GK		WKY		GK		WKY	
TF	♯	TF	♯	TF	♯	TF	♯	TF	♯	TF	♯
SP1	10	**SP1**	19	**SP1**	39	**SP1**	18	**SP1**	12	**SP1**	5
		SP3	8	**SP3**	11	HNF4A	6	**SP3**	3	**FOXO3**	3
		TP53	4	**TP53**	11	**FOXO3**	4				
				EGR1	6						
				NRF1	6						
				TCFAP2A	5						

B. MR candidates inferred by the path consistency algorithm

We first inferred six networks of all genes on the microarray for each of three periods in GK and WKY rats, by the path consistency algorithm, and then the TF-gene relationships were extracted from each network. After the extraction, then, only the relationships that include the genes with the significant difference between GK and WKY rats were further selected for 6 sets of the relationships. First, we selected the relationships in terms of specificity: TFs were extracted from the relationships that were found in GK but not in WKY. As a result, 108 TFs were identified as the MRs in Table III. The number of candidates seems large in comparison with the candidate number, 27 TFs, in the previous case of the brain tumor [5], but three networks for the three periods in GK rat were surveyed to select the candidates in the present study. The number of TFs extracted from one network, 36 TFs on average, is similar to that in the previous study.

Next, the TF-gene relationships were selected by the coverage. We selected TF-gene relationships by setting each threshold for each period in GK and WKY in Table IV. In contrast to the case by network screening, only a few TFs

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

TABLE III
TFs IDENTIFIED BY NETWORK INFERENCE IN TERMS OF SPECIFICITY.

Alx1, Arnt, Cebpg, Ddit3, Dlx5, Dmrt2, Dnmt1, Dr1, Ebf1, Elf5, Elk3, Elk4, Erg, Etv4, Etv5, Fev, Fosl1, Foxe1, Foxg1, Foxo3, Foxp4, Gabpb1l, Gfi1 Gtf2a1, Gtf2b, Gtf2e1, Gzf1, Hcfc1, Hey1, Hhex, Hoxb3, Hoxb7, Ilf3, Irx2, Kcnip4, Klf1, Klf15, Klf3, Klf5, Klf7, Ldb2, LOC680117, Mafk, Meis2, Mnat1, Msx1, Msx2, Mybl2, Myc, Myocd, Myod1, Mzf1, Neurod2, Nfix, Nfx1, Nkx6-1, Notch1, Nr1h4, Nr2f1, Nr4a1, Nr5a1, Pax8, Pbx2, Phox2a,Pitx1, Pitx3, Pou2f3, Pou3f1, Ppard, Pparg, Ppargc1a, Rbl1, RGD1566107, Rreb1, Runx1, Shh, Six5, Six6, Skp2, Sox10, Sox11, Sp1, Sp2, Spdef, Srebf1, Ss18l1, Stat5a, Stat5b, Taf2, Tbx18, Tbx2, Tcf12, Tcfap2b, Tead1, Tfdp2, Tfec, Tmf1, Tp53bp1, Twist1, Vdr, Zbtb5, Zfhx3, Zfp191, Zfp238, Zfp423, Zfp444, Zhx1, Zic1

emerged in both GK and WKY. Indeed, 44 TFs are listed in total in Table IV, and only two TFs (Tbpl1 and Cbfb) emerged in both GK and WKY. Finally we found 42 TFs that regulates 725 genes.

TABLE IV
TFs IDENTIFIED BY NETWORK INFERENCE IN TERMS OF SPECIFICITY. ♯ MEANS THE NUMBER OF REGULATED GENES. TFs FOUND IN BOTH GK AND WKY ARE INDICATED BY BOLD LETTERS.

4w				8w_12w				16w_20w			
GK		WKY		GK		WKY		GK		WKY	
TF	♯	TF	♯	TF	♯	TF	♯	TF	♯	TF	♯
Arntl	31	Max	10	Lhx5	24	Ywhae	18	Fus	10	Foxq1	32
Lhx2	22	Otx2	10	Etv1	23	Pfdn5	13	Smad5	10	Hoxa1	16
Sp2	18	Daxx	9	Ctnnb	18	Atf1	11	Nfx1	9	Rbl2	16
Gbpa	13	Sim1	9	Rpa3	8	Cdk9	11	Hsf1	8	Zic2	12
Xpa	4	Tcf21	8	Zfp105	8	Hmgb2	11	Tlx3	8	Rorc	8
Foxs1	3	Gata5	7	Foxo3	7	Sfpq	9	Tp53	8	Tcfap4	6
		Tcfap2c	7	Hoxc5	6	Zfp281	9	Foxs1	7	Pttg1	5
		Meis3	5	Litaf	6	Cdk7	8	LOC 679869	7	Ncoa3	4
		Rorc	5	Nr2f2	6	Ets2	8	**Cbfb**	6	Ccnh	3
		Snapc1	5	Foxo1	5	Hoxa1	8	Ctcf	6	Hif1a	3
		Zic2	5	Msx1	5	Nfe2l2	8	Glis2	6	Junb	3
		Meis1	4	Myocd	5	Nfil3	8	Irf7	6	Kcnip1	3
		Pou2af1	4	Pbx1	5	Six4	8	Nfkbib	6	Mtf1	3
		Srf	4	**Tbpl1**	5	Cux2	7	Nr1i2	6	Zfp148	3
		Stox2	4	Vdr	5	Mafg	7	Hdac1	5		
		Tcfcp2l1	4	Hltf	4	Nfkbia	7	Rfx5	5		
		Gtf2h2	3	Htt	4	Pgr	7	Tle1	5		
		Zfx	3	LOC 680117	4	Ppp1 r13b	7	Xpa	5		
				Mbd1	4	**Tbpl1**	7				
				Parp1	4	**Cbfb**	6				
				Rreb1	4	Ezh2	6				
				Smarcc1	4	Hbp1	6				
						Junb	6				
						Taf13	6				
						Tef	6				

C. MR selection by comparison of the TF sets detected by the two methods

We summarized the TFs detected by the two methods in terms of two criteria (Tables I-IV) in Table V. 21 TFs detected by network screening in terms of specificity overlapped with only 4 TFs and 2 TFs by path consistency algorithm by two criteria, respectively. In contrast, 3 TFs showed no overlapped TFs by path consistency algorithm by two criteria. As pointed out, one of pitfalls in network screening is the restriction of TF-gene relationships. Thus, the coverage may not be effective as a criterion in selecting TFs.

As a result, 5 TFs are finally identified as the candidates of MRs for diabetes progression in GK rats: one of 6TFs, SP2, emerged in both 4TFs and 2TFs. The TFs and their regulated genes are listed in Table VI. By preliminary survey, all of the 5TFs are not reported their any direct relations to diabetes, but related to various diseases. Notably, the relations

TABLE V
SUMMARY OF TFs.

		path consistency algorithm	
		specificity (108)	coverage (42)
network screening	specificity (21)	4	2
	coverage (3)	0	0

of Etv4 and Tcfap2b to adipogenesis, which is well known to be highly related to diabetes, are reported, together with their association of the other pathways [17], [18]. The molecular functions of the remaining three TFs, Fus, Nr2f1, and Sp2, are also reported to be related to some diseases [19]–[21]. Although direct evidence is not found in previous knowledge, 5 TFs are expected to be MR candidates, in consideration with the circumstance evidence of the relations to diseases, especially the relations of Etv4 and Tcfap2b to adipogenesis, in addition to the correct finding of new MRs in brain tumor by more preliminary procedure than that in the present study. In addition to their regulated genes, some experimental verification of MR candidates may be desirable to further examine their plausibility as the MR candidates for diabetes progression.

TABLE VI
CANDIDATES OF TF-GENE RELATIONSHIPS FOR DIABETES PROGRESSION IN GK RAT. THE GENES IN BOLD CHARACTERS ARE INCLUDED IN KNOWN TF-GENE RELATIONSHIPS DETECTED BY NETWORK SCREENING.

TF	Regulated genes					
Etv4	Mcm10	**ERBB2**	**MMP7**	**NID1**	PLAU	PTGS2
Fus	Mcpt8l2	Mcpt9	**PAICS**	**PPAT**	Ugt1a1	Ugt1a2
	Ugt1a3	Ugt1a5	Ugt1a6	Ugt1a7c	Ugt1a8	Ugt1a9
Nr2f1	ALOX5	**CPT1B**	CYP11B2	**TF**	Ugt1a3	Ugt1a5
Sp2	**CAPNS1**	**IRS2**	LOC685183	LOC685226	LOC685291	LOC685759
	LOC688519	LOC688603	LOC689083	LOC689312	LOC689338	LOC689690
	LOC689999	LC690179	LOC690328	LOC690379	LOC690577	LOC691712
	LOC691735	LOC691754	PAPSS2	Vom2r45	Vom2r46	Vom2r47
Tcfap2b	Aqp1	EGFR	KRT14	PTGDS	SOD2	TGM1

D. Remarks

In this study, we have identified the candidates of master regulators based on our previous study [5], by using an improved method based on our previous methods for identifying master regulator candidates [6]. The MR candidates were extracted from the active networks of many genes characterized by biological pathways, to present the feasible gene candidates for experimental verification. In methodological aspect, the method was improved by considering the coverage of TF by a statistical way, in addition to the specificity that considered in the previous method. At any rate, the present study illustrated one of the rational ways to narrow down the genes of MR candidates, definitely different from the metaphysical presentation such as biological pathway or large network form.

Our study intended to identify the candidates of MR, which indicated that the gene(s) had large impacts on the phenotype changes in terms of biological sense [3]. Here, we identified logically MR candidates by the specificity on TF appearance and the coverage of regulated genes to gene expression signature in the networks of GK and WKY rats.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

Apart from biological sense, we further investigate the meaning of "master" in view from the network structure. To do this, we uncovered hierarchical structures [22] of networks in 8w-12w and 16w-20w by network screening, and allocated the present 5 TFs into the hierarchical structures (Shown in Fig. 3). As seen in the figures, all of the 5 TFs were allocated into the highest level. Indeed, Nr2tf1 at 8w-12w network and Tcfap2b at 16w-20w network were definitely allocated into the highest level of hierarchical structures. Furthermore, the remaining TFs were allocated into the levels including the highest and middle levels, but not into the lowest level. In addition, previous hierarchical analysis of the regulatory networks in *Escherichia coli* and *Saccharomyces cerevisiae* suggested that MRs were in the middle of the hierarchy [23]. Although the verification experiments remain to be performed for justification of MR in terms of biological sense, present 5 TFs may be regarded as the plausible MR candidates in terms of the network structure.

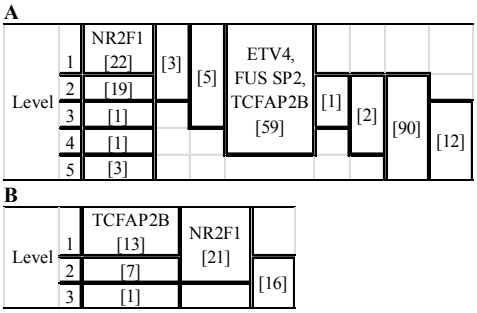

Fig. 3. Hierarchical structures of networks for 8w-12w (A) and 16w-20w (B). 5 TFs are indicated at the levels in hierarchical structures obtained by vertex-sort algorithm [22], and the numbers of TFs in each level are indicated in parenthesis.

Here we present the candidates of MRs for the diabetes progression, 5 TFs and their regulated genes in GK rat, by our original method. Fortunately, the number of candidates was very small, and may be enough to perform experiments for their verifications. Furthermore, the recent availability of the next-generation sequencer may confirm the effectiveness of our procedure, and allow further chances for testing its performance with other data sets. Indeed, RNA-seq and ChIP-seq are useful for more accurate measurements of gene expression and concrete information about the regulated genes. Thus, the combined procedure using the two approaches may compensate for the possible pitfalls of each approach, and will provide some clues about the transcriptional networks that regulate transitions into physiological or pathological cellular states of diabetes.

ACKNOWLEDGMENTS

This work was supported by a grant, "Joint Seminar 2011 in NSFC-JSPS Scientific Cooperation Program", from National Natural Science Foundation of China (NSFC) and Japan Society for the Promotion of Science (JSPS). Z.P.L. was supported by the Knowledge Innovation Program of SIBS of CAS with Grant No. 2011KIP203 and Shanghai NSF under Grant No. 11ZR1443100. This work is also partly supported by a project grant, entitled "Development of Analysis Technology for Induced Pluripotent Stem (iPS) Cell", from The New Energy and Industrial Technology Development Organization (NEDO), Japan.

REFERENCES

[1] A. A. Margolin, et al.: "Reverse engineering cellular networks", Nature Protocols, 2006, 1, 662-671.
[2] K. M. Mani, et al.: "A systems biology approach to prediction of oncogenes and perturbation targets in B cell lymphomas", Mol. Syst. Biol., 2008, 4, 169-178.
[3] M. S. Carro, et al.: "The transcriptional network for mesenchymal transformation of brain tumours", Nature, 2010, 463, 318-325.
[4] L. Chen, R.-S. Wang and X.-S. Zhang: "Biomolecular Networks: Methods and Applications in Systems Biology", Wiley, 2009.
[5] S. Saito, et al.: "Identification of Master Regulator Candidates in Conjunction with Network Screening and Inference", Int. J. Data Mining and Bioinformatics, in press.
[6] H. Zhou, et al.: "Network Screening of Goto-Kakizaki Rat Liver Microarray Data during Diabetic Progression", BMC Sys. Biol., 2011, 5(Suppl 1), S16.
[7] S. Saito, et al.: "Potential linkages between the inner and outer cellular states of human induced pluripotent stem cells", BMC Sys. Biol., 2011, 5(Suppl 1), S17.
[8] E. Wingender: "TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation", Brief. Bioinformatics, 2008, 9, 326-332.
[9] A. Subramanian, et al.: "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles", Proc. Natl. Acad. Sci. USA, 2005, 102, 15545-15550.
[10] S. Saito, et al.: "Network evaluation from the consistency of the graph structure with the measured data", BMC Sys. Biol., 2008, 2, 84.
[11] P. Spirtes, C. Glymour and R. Scheines: "Causation, Prediction, and Search" (Springer Lecture Notes in Statistics, 2nd edition, revised), MIT Press, Cambridge, 2001.
[12] S. Saito and K. Horimoto: "Co-Expressed Gene Assessment Based on the Path Consistency Algorithm: Operon Detention in Escherichia coli", Proceedings of IEEE International Conference on Systems, Man and Cybernetics, 2009, pp. 4280-4286.
[13] S. Saito, et al.: "Discovery of Chemical Compound Groups with Common Structures by a Network Analysis Approach", J. Chem. Inf. Model., 2011, 51, 61-68.
[14] Y. Benjamini and D. Yekutieli: "The control of the false discovery rate in multiple testing under dependency", Annals of Statistics, 2001, 29, 1165-1188.
[15] F.E. Grubbs: "Sample criteria for testing outlying observations", Ann. Math. Statistics, 1950, 21, 27-58.
[16] R.P. Almon, D.C. DuBois, W. Lai, B. Xue, J. Nie, W.J. Jusko: "Gene expression analysis of hepatic roles in cause and development of diabetes in Goto-Kakizaki rats", J. Endocrinology, 2009, 200, 331-346.
[17] K. W. Park, et al.: "The small molecule phenamil is a modulator of adipocyte differentiation and PPARγ expression", J. Lipid Res., 2010, 51, 2775-2784.
[18] Y. Tao, et al.: "The transcription factor AP-2beta causes cell enlargement and insulin resistance in 3T3-L1 adipocytes", Endocrinology, 2006, 147, 1685-1696.
[19] T. J. Jr Kwiatkowski, et al.: "Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis", Science, 2009, 323, 1205-1208.
[20] K. K. Brown, et al.: "NR2F1 deletion in a patient with a de novo paracentric inversion, inv(5)(q15q33.2), and syndromic deafness", American journal of medical genetics Part A, 2009, 149A, 931-938.
[21] M. Letourneur, et al.: "Sp2 regulates interferon-gamma-mediated socs1 gene expression", Mol. Immunol., 2009, 46, 2151-2160.
[22] R. Jothi, et al.: "Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture:, Mol. Sys. Biol., 2009, 5, 294.
[23] H. Yu and M. Gerstein: "Genomic analysis of the hierarchical structure of regulatory networks", Proc. Natl. Acad. Sci. USA, 2006, 103, 14724-14731.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE