The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Cross-species Functional Conservation and *in silico* Gene Replacement for HIF Pathways

Chia-Sheng Chuang[1] Chih-Chia Chen[1] Tun-Wen Pai[1]

Chin-Hua Hu[2] Hao-Teng Chang[3] Margaret Dah-Tsyr Chang[4]

[1]Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan
[2]Institute of Bioscience and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan
[3]Institute of Molecular Systems Biomedicine, China Medical University, Taichung, Taiwan
[4]Institute of Molecular and Cellular Biology, National Tsing Hua University, Hsinchu, Taiwan

Abstract Hypoxia-inducible factors (HIFs) are transcription factors that play a crucial role in adaptive processes to a hypoxic cellular environment for most of organisms. The activation of HIF pathway has significant impacts on gene expression patterns in cancer research and it is important to analyze functional conservation and possible gene replacement within the HIF pathway for cancer therapies. Phylogenomics is a useful tool to identify the relationship among various species. By evaluating the evidence of homologous relationship within cross-species pathways, functional conservation and gene replacement for a specific biological function can be obtained. Mapping orthologous and paralogous genes through a phylogenetic tree onto the conserved pathways indicates a strong homology of function. This study retrieved an initial pathway from the KEGG pathway database, and constructed an ontology table through cross-species comparison. A quantitatively measured HIF pathway was depicted to illustrate suitable simulations of cellular function, and the built-up map discovered all substitutable genes and explored unknown subpathways by employing homologous appearing rates. Novel terminologies of OrthRate and ParaRate are proposed to quantitatively indicate the flexibility of a homologous pathway and to enhance the substitutable possibilities of functional genes. This is the first novel system which can generate homologous biological pathways for various organisms based on orthologous and paralogous gene analyses.

Keywords HIF (Hypoxia-inducible factor); KEGG (Kyoto Encyclopedia of Genes and Genomes); Orthology; Paralogy; Signaling pathway

1 Introduction

Hypoxia-inducible factors (HIFs) are transcription factors that play a crucial role in adaptive processes to a hypoxic cellular environment [1-2]. These responses include pulmonary vascular constriction to redirect blood flow and erythropoietin secretion to boost red blood cell production. There are three known isoforms of HIF family (HIF-1, HIF-2 and HIF-3) and HIF is a heterodimeric protein composed of

an oxygen-sensitive α-subunit in cytoplasm and a constitutively expressed β-subunit in nucleus[3]. HIFs facilitate heterodimerization of HIF-α with HIF-β through the Per-ARNT-Sim (PAS) domain and control target genes by DNA binding through basic-helix-loop-helix (bHLH) domain [4-5]. The transcriptional activity and formation of HIF heterodimer depend on the stability of HIF-α which is regulated by oxygen tension. In fact, the pressure of cellular oxygen is sensed by prolyl-hydroxylase domain enzyme family which contains three members: PHD1, PHD2 and PHD3. Under hypoxic conditions, PHDs decrease hydroxylation ability so that HIF-α accumulates [6]. Not only PHDs but many proteins can interact with HIF-α. The regulation mechanisms are achieved by modification such as ubiquitination, acetylation, hydroxylation or phosphorylation, once stabilized HIFs interact with co-activators such as P300/CBP and then bind to a hypoxia response element (HRE) in regulatory regions of hypoxia-inducible genes. The actual HIF DNA binding sites within an HRE are referred to as a core HRE, 5'- RCGTG – 3' (where R is A or G) [7]. The consensus HRE motif is useful for identifying novel HIF target genes, as well as referred to as a *cis*-regulatory element required for gene transcription in hypoxia signaling pathway [8-9]. Recently, it is shown that the inhibition of HIF-1 activity may be a strategy in new cancer therapy [10].

A metabolic pathway represents a sequence of chemical reactions occurring within a cell usually controlled and catalyzed by enzymes, and by which one organic substance is converted to another. Metabolic pathways almost keep the same function working in any growth stage to maintain cell living but signaling pathways do not. Signaling pathways (also called signal transduction) start with a signal to a receptor, and end with a change in cell behaviors [11]. In other words, signal pathway has been implicated in different stages of dynamic transitions of a network. Although there are some differences between them, both types of pathways are important components of system biology and often work together to carry on physiology mechanism [12]. Taking the HIF pathway as an example, it is identified as a signaling pathway under low oxygen supply [13-14] and HIF stability is regulated by PHD and FIH (Factor-Inhibiting HIF) [15]. The classical representation of a biological pathway provides varied associations among genes and proteins. Besides, a biological pathway can provide system-level insight to discover functional information through interacting relationship. During the last decade, an increasing number of pathway datasets have been established in order to combine functions and interactions within a network representation and to discover cell behaviors. Hence, analyzing and understanding a biological pathway can facilitate biologists to design experiments through various aspects and approaches.

In taxonomy, phylogenomics indicates the construction of a phylogeny based on evolutionary conservation of genomes of related organisms. Accordingly, the phylogenetic analysis validly provides evolutionary relationships among various species [16], and the phylogenetic inferences can facilitate understanding the derivation of species and delineating homologous genes [17]. Homologous genes are evolutionary results of random mutations accumulated over the course of many generations, which are composed of two major categories: orthologous and paralogous genes. Orthologous genes are evaluated directly from an ancestral gene through speciation events to daughter species and paralogous genes are diverged

after duplication events within one species. Furthermore, most orthologous genes retain similar function in the course of evolution, while paralogous gene may evolve new functions even if these genes are related to the original one. It is well known that the history of orthologous genes reflect the development history of diversified species [18]. Up to date, there are several ways to integrate biological information from different gene families to form a single species phylogeny, e.g. integrating multiple gene family trees to form a single phylogenetic tree. Determining such evolutionary relationships from phylogenetic analysis provides several advantages including facilitating the prediction of protein functions and the identification of horizontal gene transfer events [19]. In this study, we have proposed a mechanism to predict possible substituent genes from related homologous genes which satisfy requirements of corresponding biological functions. If a substituent gene possess similar function and is able to replace the original one, a novel pathway may be discovered in future research. Here, the enumerated subject focused on HIF responses and the HIF biological pathways from several well known phylogenetic trees were adopted and illustrated in this study. Biological function of a living cell is a result of many interacting molecules, and it cannot be simply attributed to a single gene or a signal molecule. It is important that a biological pathway can comprehensively represent higher order functions in terms of the network of interacting molecules. Therefore, the main goal of this study is to construct conserved pathways through cross-species comparison and orthologous gene clustering operations. More importantly, when a partial subset of the constructed pathway of the target species disappears, the proposed mechanism can efficiently replace the empty nodes by retrieval of corresponding paralogous genes, and those genes may possibly possess similar biological functions [20]. In this study, we have successfully illustrated an example to reveal and verify functional conservation and gene replaceability of HIF pathways among various species.

2 Materials and Methods

In this study, we have collected and integrated several well known representative datasets for verifying the annotated HIF biological pathways, including corresponding information of homologous genes. First, the initial pathway information was retrieved from KEGG (Kyoto Encyclopedia of Genes and Genomes) database. KEGG is a knowledge database for systematic analysis of gene functions, linking genomic information with higher order functional information [21-22]. Each map or pathway in KEGG is categorized into an existing taxonomy according to its function, and each pathway is supplemented with a set of orthologously grouped tables for the cross-species information with respect to conserved pathways. The orthologous table summaries functional correlations in the pathway, physical correlations in genomes, and evolutionary relationships among species. It provides useful information as a reference dataset for functional annotations. Taking HIF pathway as an example, from the main interface of KEGG (http://www.genome.jp/kegg/pathway.html/), if the user enters the keywords “hypoxia inducible factor”, and the system responses with only one entry of map05211. In addition to the searched pathways, users can adopt orthologous

information from KEGG and/or Ensembl databases. Especially, orthologous genes identified in KEGG are not only achieved by evaluating sequence similarity but also examining if all constituent members are verified within a functional group, such as a conserved subpathway or a molecular complex. The variation of derived datasets from these two resources may provide a hint for a possible modification of an accurate pathway. To compare the expression level of orthologous gene cluster in pathways against others by utilizing a quantitative measure, it is necessary to represent those associated genes as mathematical objects and provide measurable indices for effective representation. Here, we define two types of homologous rate (OrthRate & ParaRate) that can quantitatively characterize the substitutable possibilities for functional genes. The OrthRate is defined as the total number of corresponding orthologous genes within a specified pathway from the query specie divided by the total number of associated genes within the identical pathway from the target species. The OrthRate stands for the proportional percentage of corresponding genes within cross-species biological pathways. The ParaRate is defined as a replacement ratio of the number of paralogous genes from the target gene to those of corresponding orthologous genes. According to the statistic results, the system can be expected to retrieve possible missed subpathways within an individual species and to predict extra direct and/or indirect pathways within each species. To demonstrate the functional conservation and replaceability of genes among various species, only 6 remote model species were considered for orthologous analysis in this study, including *Ciona intestinalis* (CIN), *Xenopus tropicalis* (XTR), *Gallus gallus* (GGA), *Mus musculus* (MMU), *Danio rerio* (DRE) and *Homo sapiens* (HSA).

Biological network analysis provides valuable insight into protein and genetic interactions for model species. In general, a biological network consists of a number of molecular components and is fraught with a wide range of statistical, mathematical, and computational related issues. Here, the HIF pathway statistics was illustrated and designed to capture the characteristics of hypoxia network**.** To demonstrate the validity of *in silico* pathways in biological sense, existence of binding sites within HIF target genes are considered as the verification criteria. HIFs bind target genes at the functional hypoxia response elements (HREs). An overview of the known target genes of HIF revealed that the length of a HRE is nearly 18 base pairs. The mandatory core HRE sequence is “CGTG”, and it is the minimal DNA motif required for interacting with HIF. The appearance frequency of HREs located within flanking sequence is randomly appeared as previous reports [7]. To perform aligning processes and identify whether the HREs are located within the paralogous genes, we have employed position-specific scoring matrices (PSSM) to extract all HRE candidates within those retrieved HIF target genes. The PSSM matching mechanism efficiently scans through a DNA sequence and identifies the most probable motifs with precise locations [23-24]. The frequency profile of HREs for HIF target genes is generated according to the published paper by R.H. Wenger, *et al*. [7]. The PSSM based motif searching mechanism allows us to verify functional conservation of HIF responses of paralogous genes. In this study, identified HRE motifs are listed according to their ranking order based on matched scores. The top 3 matched motifs were listed for illustration, nevertheless,

more than 3 HRE motifs within a paralogous gene were very likely for all the retrieved HIF target genes.

To understand the hypoxia signal transduction mechanism, the approved HIF pathway from KEGG was obtained initially. The KEGG database provides the information of all related genes participating in HIF pathway. By calculating previously defined ParaRate, the proposed system can discover all novel substituent genes that may possess strong possibilities to replace the original one without losing its biological function. According to the obtained OrthRate percentages, the system derives similar gene cluster through cross-species comparisons and predicts the functional pathway for query species. Finally, comparing the paralogous genes obtained from HIF target genes facilitates the discovery of probable substituent genes within the biological pathway. Accordingly, it is possible to construct a novel functional subpathway with respect to these replaceable genes.

3 Result and Discussion

Different signaling pathways sometimes possess crosstalk phenomena within a cell. For example, the HIF responses to a signal inducing condition activating multiple responses in a cell or an organism. Based on the evidence of evolutionary conservation, we have made assumptions that predicting a novel biological pathway of a specified function of a designated species by observing cross-species conservation is achievable. Through comparing orthologous relationships among various species and paralogous properties within an individual species, orthologous relationships shows the conservation of genes in a strong sense which does not only hold reservation in species evolution but also reflect importance in functional pathways. Taking HIF target genes as an example as shown in Table 1, the OrthNum (the number of orthologous genes) of HSA, MMU, GGA and DRE possess almost equivalent number of genes in this network. In contrast to those species, XTR and CIN possess less numbers of genes than the others. This result implies that the former four species possess more similar mechanism to maintain the HIF function than the last two species. We can easily determine the functional similarity by index of OrthRate, and further provide a tip to suggest one of suitable candidates of model organism for performing subsequent biological experiments. In this proposed system, a novel subpathway with substituent genes is constructed according to the characteristics of paralogous gene distribution. A biological network with higher flexibility enhances the survival rates when patients with cancer are taking clinical drug therapy, which is caused by turning on the novel pathway to maintain original physiology mechanism normally. Hence, the ParaRate is applied to manifest the flexibility and replacement. On the other hand, a phylogenetic tree shows DRE possesses a longer evolution distance to HSA than MMU. However, both DRE and MMU hold high ParaRates as HSA to show the strength and flexibility of the pathway regarding to evolution processes. It is important that the similarity of HIF pathways between HSA and DRE is beyond our expectations. Therefore, predicting novel paralogous genes is a good starting point to obtain a flexible and substitutable pathway.

Table 1: Phylogeny shows the evolution distance from HSA. OrthRate represents the functional conservation within the HIF pathway. ParaRate represents the relative possibility of gene replacement.

	Phylogeny	OrthNum	OrthRate	ParaNum	ParaRate
HSA	1	14	N/A	24	N/A
MMU	2	14	14/14(100%)	26	26/24(108.3%)
GGA	3	13	13/14(92.9%)	20	20/24(83%)
XTR	4	9	9/14(64.3%)	8	8/24(33%)
DRE	5	12	12/14(85.7%)	23	23/24(95.8%)
CIN	6	9	9/14(64.3%)	9	9/24(37.5%)

Based on the HIF pathway (map05211) from KEGG, we have collected HIF target genes and their corresponding paralogous genes. According to previous related literatures, functional HREs appear within promoter regions. Hence we have tried to identify whether the retrieved paralogous genes also possessing HREs in the same regions. Employing PSSM alignment algorithms, the exact core HRE motifs and the approximate segments on both sides of HRE core segment were obtained. For example, homologous genes of VEGFA_B (HSA:3, MMU:3, GGA:1, DRE:1) are shown in Table 2. Here we only list the top 3 matched motifs and their corresponding locations precisely. Obviously, both XTR and CIN do not hold the important HRE motifs as a target gene of HIFs. Besides, the results demonstrated that paralogous genes also possess HRE motifs and can be considered as a good substituent gene for the original gene intuitively. There might exist a lot of genes possessing similar HRE motif patterns. However, these genes are lack of paralogous relationship in our study.

Table 2: Detected HREs within upstream 2000 base pairs from homologous genes of K05448 (VEGFA_B).

HSA	ENSG00000119630 GGCAGG**CGTG**CAGACTCA Loc:-963~-946 TGTGTC**CGTG**CCTGGCTA Loc:-1392~-1375 GCCCCT**CGTG**GGTGGGCA Loc:-628~-611	ENSG00000112715 TGAGGA**CGTG**TGTGTCTG Loc:-517~-500 TGCATA**CGTG**GGCTCCAA Loc:-982~-965 TGTGTG**CGTG**TGGGGTTG Loc:-485~-468	ENSG00000173511 CGAGAT**CGTG**CCCCGGGG Loc:-641~-624 GGAGCG**CGTG**TCTGGGTC Loc:-277~-260 CTCACG**CGTG**CCACGGAG Loc:-1601~-1584
MMU	ENSMUSG00000004791 TGAGCA**CGTG**TGGATCCT Loc:-542~-525 CCAATC**CGTG**TGTGCTCA Loc:-204~-187 ATGTCA**CGTG**AAATGACG Loc:-122~-105	ENSMUSG00000023951 TGCATA**CGTG**GGTTTCCA Loc:-1082~-1065 AGTCTG**CGTG**AGGGAGGA Loc:-1538~-1521 TGAGTG**CGTG**CATGCATG Loc:-1570~-1553	ENSMUSG00000024962 TCCCCT**CGTG**AGGCAGCG Loc:-1799~-1782 ACTACA**CGTG**CAATAAAC Loc:-1726~-1709 GTCAAG**CGTG**CTGAGGCC Loc:-287~-270
GGA	ENSGALG00000010290 CCCCGA**CGTG**CGGAGCGG Loc:-1970~-1953 TGGCAC**CGTG**CTGGAATA Loc:-143~-126 CCCCAT**CGTG**CAGCCCCA Loc:-208~-191	N/A	N/A

DRE	ENSDARG00000034700 CCTGTA**CGTG**GTGATGGA Loc:-997~-980 TATCGT**CGTG**TTGTGATT Loc:-1106~-1089 TTAAAC**CGTG**TGCGCTGC Loc:-55~-38	N/A	N/A

The results of our research present a new paradigm for investigating functional diversification. A gene performs different biological function when it participates in different functional pathway. Quantitative analysis of biological pathway suggests that function of paralogous transcription factors might be inconsistent with evolutionary expectation. In Figure 1, each node in the pathway denotes the exact number of paralogous genes and the orthologous property between HSA and DRE. Black background squares indicate that the DRE species may hold higher flexibility and possibility than HSA regarding to the gene substitution. Gray background squares mean equivalent number, while the white ones indicate that DRE do not hold any correlated paralogous gene in this pathway.

When a biological pathway has been described for an organism, may it be applied for presenting another specified species? Which gene can be involved within this pathway? To enhance the functional conservation among species, we tried to extend pathways by homology relationship and to discover the clues in this research. The *in silico* predicted biological pathways under hypothetical assumptions may be not persuasive and authentic. However, it can efficiently provide a clue to explain the functional conservation and gene replacement for cross-species pathways.

Taking the analysis of hypoxia responses as an example, it is currently difficult to estimate how many HIF target genes exist in an organism. This study serves as a pioneer guide to discover paralogous genes with HRE motifs. Currently, our analysis only focuses on six species and it is not strong enough to explain whole possible pathway models. However, the analytical mechanism for extra model species for discovering conceal or undefined substitutable pathways can be expected in near future.

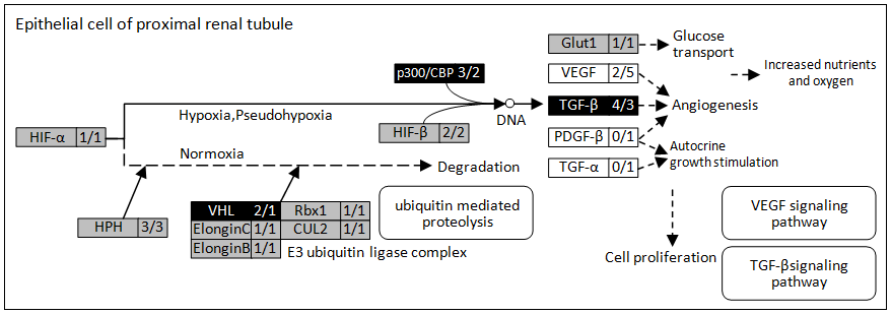

Figure 1: The Pararates of conserved genes in the HIF pathway between HSA and DRE. Numbers of paralogous genes are labeled at each node.

Indeed, the phylogenetic analysis just allows the determination of gene orthology relationships but not orthologous functions. However, we can extent to observe at whole functional pathway, not on an individual gene. It is possible that not only genes but also pathways are conserved in general evolution processes. To conclude, if this research can provide effective hints for biologists that functional paralogous genes indeed appear within the HIF biological pathway, we may be able to extend this functional conservation mechanism to all other biological pathways. Furthermore, by analyzing the conserved pathways through cross species comparison, the diversity of pathways among model species can be discovered and distinguished. Accordingly, the compared results can suggest a better model species for subsequent experiments, and it is practicable and effective for clinical researches on disease therapy and powerful tool for understanding the physiology mechanism.

Acknowledges

This work is supported by the Center for Marine Bioenvironment and Biotechnology (CMBB) in National Taiwan Ocean University, Keelung, Taiwan, and the National Science Council in Taiwan, R. O. C. (NSC98-2627-B-019-003 and NSC98-2221-E-019-031-MY2 to Tun-Wen Pai; NSC98-2627-B-007-017 and NSC99-2627-B-007-017 to Margaret D.-T. Chang).

References

[1] G.L. Semenza, *Life with oxygen.* Science, 2007, 318(5847): 62-4.
[2] R.H. Wenger, *Cellular adaptation to hypoxia: O2-sensing protein hydroxylases, hypoxia-inducible transcription factors, and O2-regulated gene expression.* FASEB J, 2002, 16(10): 1151-62.
[3] P.C. Mahon, K. Hirota, and G.L. Semenza, *FIH-1: a novel protein that interacts with HIF-1alpha and VHL to mediate repression of HIF-1 transcriptional activity.* Genes Dev, 2001, 15(20): 2675-86.
[4] Q. Ke and M. Costa, *Hypoxia-inducible factor-1 (HIF-1).* Mol Pharmacol, 2006, 70(5): 1469-80.
[5] B.H. Jiang, E. Rue, G.L. Wang, R. Roe, and G.L. Semenza, *Dimerization, DNA binding, and transactivation properties of hypoxia-inducible factor 1.* J Biol Chem, 1996, 271(30): 17771-8.
[6] A. Giatromanolaki, M.I. Koukourakis, F. Pezzella, H. Turley, E. Sivridis, D. Bouros, G. Bougioukas, A.L. Harris, and K.C. Gatter, *Expression of prolyl-hydroxylases PHD-1, 2 and 3 and of the asparagine hydroxylase FIH in non-small cell lung cancer relates to an activated HIF pathway.* Cancer Lett, 2008, 261(1): 87-93.
[7] R.H. Wenger, D.P. Stiehl, and G. Camenisch, *Integration of oxygen signaling at the consensus HRE.* Sci STKE, 2005, 2005(306): re12.
[8] Y.S. Chun, M.S. Kim, and J.W. Park, *Oxygen-dependent and -independent regulation of HIF-1alpha.* J Korean Med Sci, 2002, 17(5): 581-8.
[9] M. Safran and W.G. Kaelin, Jr., *HIF hydroxylation and the mammalian oxygen-sensing pathway.* J Clin Invest, 2003, 111(6): 779-83.
[10] E.J. Yeo, Y.S. Chun, and J.W. Park, *New anticancer strategies targeting HIF-1.* Biochem Pharmacol, 2004, 68(6): 1061-9.
[11] N.R. Gough, E.M. Adler, and J.F. Foley, *Cell Signaling: Details, Details, Details.* Sci. STKE, 2007, 2007(407): eg9-.

[12] A. Cakmak and G. Ozsoyoglu, *Mining biological networks for unknown pathways.* Bioinformatics, 2007, 23(20): 2775-83.
[13] G. Semenza, *Signal transduction to hypoxia-inducible factor 1*. Biochem Pharmacol, 2002, 64(5-6): 993-8.
[14] G.L. Semenza, *Hypoxia-inducible factor 1 (HIF-1) pathway*. Sci STKE, 2007, 2007(407): cm8.
[15] F. Dayan, D. Roux, M.C. Brahimi-Horn, J. Pouyssegur, and N.M. Mazure, *The oxygen sensor factor-inhibiting hypoxia-inducible factor-1 controls expression of distinct genes through the bifunctional transcriptional character of hypoxia-inducible factor-1alpha.* Cancer Res, 2006, 66(7): 3688-98.
[16]H. Philippe, R. Derelle, P. Lopez, K. Pick, C. Borchiellini, N. Boury-Esnault, J. Vacelet, E. Renard, E. Houliston, E. Queinnec, C. Da Silva, P. Wincker, H. Le Guyader, S. Leys, D.J. Jackson, F. Schreiber, D. Erpenbeck, B. Morgenstern, G. Worheide, and M. Manuel, *Phylogenomics revives traditional views on deep animal relationships.* Curr Biol, 2009, 19(8): 706-12.
[17] R.L. Tatusov, D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin, *The COG database: new developments in phylogenetic classification of proteins from complete genomes.* Nucleic Acids Res, 2001, 29(1): 22-8.
[18] T. Tsuru, M. Kawai, Y. Mizutani-Ui, I. Uchiyama, and I. Kobayashi, *Evolution of paralogous genes: Reconstruction of genome rearrangements through comparison of multiple genomes within Staphylococcus aureus.* Mol Biol Evol, 2006, 23(6): 1269-85.
[19] J. Raes and Y. Van de Peer, *Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico.* Appl Bioinformatics, 2003, 2(2): 91-101.
[20] N. Yosef, R. Sharan, and W.S. Noble, *Improved network-based identification of protein orthologs.* Bioinformatics, 2008, 24(16): i200-6.
[21] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, *KEGG for linking genomes to life and the environment.* Nucleic Acids Res, 2008, 36(Database issue): D480-4.
[22] J. Wixon and D. Kell, *The Kyoto encyclopedia of genes and genomes--KEGG.* Yeast, 2000, 17(1): 48-55.
[23] N. Sen, M. Mishra, F. Khan, A. Meena, and A. Sharma, *D-MATRIX: A web tool for constructing weight matrix of conserved DNA motifs.* Bioinformation, 2009, 3(10): 415-8.
[24] C.T. Workman, Y. Yin, D.L. Corcoran, T. Ideker, G.D. Stormo, and P.V. Benos, *enoLOGOS: a versatile web tool for energy normalized sequence logos.* Nucleic Acids Res, 2005, 33(Web Server issue): W389-92.