Differences in Promoters of Orthologous Genes

Youlian Pan*[,1], Sieu Phan[1], Fazel Famili[1], Maria Luz Jaramillo[2], Anne Engelina Gesina Lenferink[2] and Edwin Wang[2]

[1]Institute for Information Technology, NRC, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada; [2]Biotechnology Research Institute, NRC, 6100 Royalmount Ave., Montreal, Quebec, H4P 2R2, Canada

Abstract: Human genetic experiments are often conducted based on the orthologous genes in other mammals such as mouse and rat. The resulting conclusions of such experiments are often limited in their applicability to the human situation. This has raised a question as to why the orthologous genes with closely related or even identical coding regions behave differently in various mammals, and motivated us to study the promoter of these genes. We proposed a functional promoter similarity index (FPSI) based on the number of putative, but statistically significant associations ($p \leq 0.05$) between transcription factors and their target orthologous genes. We deduced such association through searching known transcription factor binding sites from promoters of the genes. The FPSI was validated using microarray gene expression data. We did pair-wise study of seven vertebrate genomes (human, chimpanzee, mouse, rat, dog, chicken, and zebrafish). The FPSIs of orthologous genes are generally high between human and chimpanzee, with a mean FPSI of 0.79, but gradually decrease when human is compared to the mouse (0.22), rat (0.2), dog (0.2), chicken (0.13) or zebrafish (0.06). We then performed an analogous analysis for 2128 human cancer-associated genes and the results were similar, but had significantly improved FPSIs between these human genes and their orthologs in mouse, rat, and dog. The differences in the promoter regions of orthologous genes appear to be genome wide and negatively correlated with divergence time of the organisms. Such correlation suggests that the FPSI could be used as a measure of phylogenetic conservation.

Keywords: Orthologs, Promoter Similarity, Transcription Factor, *cis*-Regulatory Element, Conservation.

INTRODUCTION

Analysis of transcriptional regulation of a gene is one of the greatest challenges faced by researchers both in biology and in computational sciences. The availability of genomic sequences in public databases, such as the University of California Santa Cruz (UCSC) genome browser [1] and Ensembl genome browser [2], allows for the prediction of *cis*-regulatory elements in the promoter region of a gene which gives a glimpse of its transcriptional regulation. Tremendous efforts have been made in this area by many laboratories around the world and numerous computational tools have been developed for the identification of *cis*-elements and their binding transcription factors (TFs) over the past decades (see reviews [3-5]). In addition, multiple *cis*-elements that interact with the same TF have been identified through biological experiments. Based on the alignment of these *cis*-elements, a consensus motif and a positional weight matrix (PWM) can be constructed for each TF. These *cis*-elements and PWMs are available in public databases, such as TRANSFAC [6] and JASPAR [7] and can be used to search for putative TF binding sites (TFBSs) by PWM-based methods, such as Hidden Markov Model and others as reviewed in [3-5].

Homologous genes are derived from a common ancestral gene. Two classes of homology can be defined according to the mode in which these genes have diverged from their last common ancestor [8, 9]. The first class consists of the orthologs, which have diverged through speciation, whereas the second class, the paralogs, resulted from sequence duplication within the same genome. Nevertheless the two classes cannot be totally separated since paralogs can give rise to orthologs through subsequent speciation [10, 11]. In addition, a given gene in one genome can have one or more orthologs in another genome. Further details of the subcategory terminologies of orthologs and paralogs are reviewed in [12]. Nonetheless, it should be pointed out that in the genomics community ‘the same gene in different species’ is referred as orthologous genes [13].

Earlier studies on mutations mostly focused on the coding region. For example, single nucleotide polymorphisms (SNPs), which are single nucleotide mutations within the coding sequence of a gene, or, as was discovered more recently through the HapMap project, in the intergenic regions [14]. The focus has now gradually shifted to considering whether *cis*-regulatory and coding mutations make different contributions to the phenotypic difference. Several cases suggest that some phenotypic changes are more likely to have resulted from *cis*-regulatory mutations than from mutations in the coding regions of a gene [15, 16].

Cancer is a prevalent clinical problem in modern society, which is characterized by uncontrollable cell growth, evasion of death, immortality and the ability to invade and avoid detection. The American Cancer Society estimated about 1,529,560 new cancer cases are expected to be diagnosed and over 569,490 deaths in 2010 [17], whereas Canadian Cancer Statistics estimated 173,800 new cancer cases and

*Address correspondence to this author at the Institute for Information Technology, NRC, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada; Tel: 1-613-993-0853; Fax: 1-613-952-0215;
E-mail: youlian.pan@nrc-cnrc.gc.ca

76,200 deaths in 2010 [18]. It is well established that mutations in specific genes have been associated with the neoplastic transformation and development of specific cancer types [19, 20]. Coding regions of many human cancer-associated genes (CAGs) are largely identical to their orthologs in mouse and other mammalian organisms. Cancer genes are on average more conserved than other genes [21, 22]. Genes that contribute to cancer fusion are also more conserved [23]. Similarly, the essential genes are more conserved than the nonessential genes are [24].

In modern laboratories studying human diseases, metabolic functions, and genetics, experiments are often conducted on the mouse and rat models. However, the resulting conclusions of such experiments are often limited in their application to humans [25]. This has raised a question why the orthologous genes that suppose to perform the same function behave so differently in different mammals, and prompted us to study the *cis*-regulatory elements of orthologous genes.

We conducted a genome-wide pair-wise comparison between promoters of orthologous genes in human, chimpanzee, mouse, rat, dog, chicken, and zebrafish. In the following sections, we present the algorithm measuring the promoter similarity, a brief description of the methods that were used in this study, result of genome-wide comparison, discussion, and finally conclusions.

MATERIALS AND METHODOLOGY

We searched the highly significant putative TFBSs from promoter regions of all genes based on the PWMs obtained from the TRANSFAC database [6] using Profile Hidden Markov Model (PHMM) [26]. The PHMM is a well established method for sequence motif search. However, the PHMM has its drawback of selecting a threshold for significant motifs. We therefore conducted a comparative study using several PWMs for humans and yeast and proposed a threshold selection criterion based on single sequence in one of our earlier papers [27]. In this paper, we selected a stringent threshold ($p \leq 0.05$) for each of the 573 PWMs. We then searched the highly significant putative TFBSs from the promoter regions. We deduced the TF-gene associations based on these TFBSs and then compared them between the orthologous genes.

Promoter Similarity Measure Between Orthologs

A putative association between a TF and a gene is deduced based on the existence of a significant putative binding site (BS) of the TF in the promoter region of the gene. Let χ and ψ be the sets of distinct TFs that are found to be associated with the two genes *X* and *Y*, respectively. Usually, χ and ψ share some common instances. Let *n*(*Z*) be the number of TFs in a set *Z*, the promoter similarity, *Sim(X, Y)*, between two genes *X* and *Y* can be defined as:

$$Sim(X,Y) = \frac{n(\chi \cap \psi)}{n(\chi \cup \psi)}, \qquad (1)$$

Since $\chi \cap \psi$ is a subset of $\chi \cup \psi$ and $n(\chi \cap \psi) \leq n(\chi \cup \psi)$, the value of *Sim(X, Y)* is between 0 and 1. *Sim(X, Y)* is defined in such a way that the influence of the abundance of a specific TFBS in certain pairs of promoters of orthologous genes is mitigated; that is no matter how many copies of a TFBS appear on the promoter of a gene, as long as they collectively qualify the threshold of significance measure ($p \leq 0.05$), we consider one association based on Equation (1). Because this method measures promoter's functional, rather than sequential, similarity, we call it Functional Promoter Similarity Index (FPSI). This approach has been successfully applied in one of our recent studies [28].

Data Sources

Promoter sequences (1000 bp upstream and 200 bp downstream of the Transcription Start Site, RefSeq tables) of human (version=hg18), chimpanzee (version=panTro2), mouse (version=mm9), rat (version=rn4), dog (version= canFam2), chicken (version= galGal3) and zebrafish (version= danRer5) were obtained from UCSC Genome Browser [1] on May 29, 2009.

Although some promoter elements may lie a few tens of thousands bp upstream of transcription start site (TSS, see [5] and refs therein), the majority of TFBSs are usually more concentrated in the first 1000 bp or even in a closer proximity of TSS [5, 29]. Because of the fact that we require a statistical significance level of $p \leq 0.05$, all motifs of length 7 bp or below are not satisfied (see [27] for details). We included a small proportion of downstream sequence to account for alternative splicing, which result in different TSSs for the same genes and is obvious in the RefSeq tables. The difference in TSS for the same gene is believed to be within 200 bp [29].

The orthologous genes were obtained from the NCBI Homologene Database [30] (build 63, May 21, 2009). PWMs were obtained from the TRANSFAC professional database [6], release 2009.1. We retrieved 2128 human CAGs from [31].

Experiments

Promoter regions of the orthologous genes were compared and their similarity was calculated based on Equation (1). We first performed a validation of FPSI (Equation 1) by using microarray data on orthologous gene expression of lung adenocarcinoma between human and mouse [32] obtained from the ArrayExpress Gene Expression Atlas [33]. We then performed a pair-wise comparison of the seven organisms for all genes listed in the RefSeq tables and labelled it as the genome-wide promoter comparison. For genes with multiple paralogs in certain organism, we selected the best similarity between the orthologs. For example, species *A* and *B* have *n* and *m* paralogs, respectively; we did $n \times m$ pair-wise comparisons and the best FPSI among the $n \times m$ pairs is selected to represent the similarity between the orthologs. We then retrieved the FPSIs of the 2128 human CAGs from the genome-wide dataset. The distributions of the human CAGs were investigated over the functional promoter similarity profile by dividing the similarity profile into bins of 0.1 FPSI span.

RESULTS

Validation of the Promoter Similarity Index

We divided the microarray gene expression data on orthologs of lung adenocarcinoma [32] into two groups.

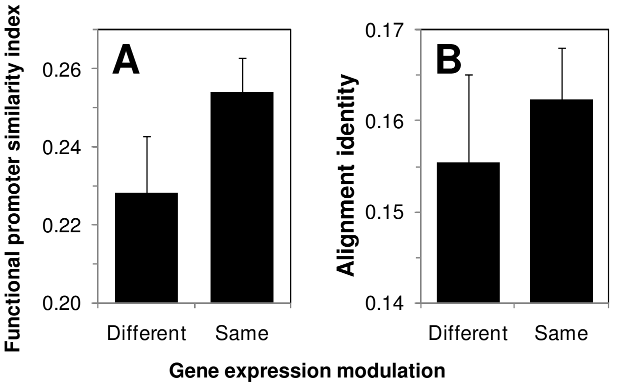

Fig. (1). Validation of the FPSI. Error bar = standard error.

Group I consists of genes that have the same gene expression modulation; both are up- or down-modulated in the orthologous genes. Group II consists of genes that have different gene expression modulation between the orthologs. We calculate promoter similarity by using Equation (1) and alignment identity using BLASTN between the entire 1200 bp promoter sequences of the orthologs of human and mouse. The result indicates that the FPSIs of Group I genes are significantly higher than those of Group II (Fig. **1**). Nevertheless, these two groups cannot be significantly separated by the identity in sequence alignment. Furthermore, we did comparison between the FPSIs on one hand and sequence conservation in the promoters and in the coding sequences on the other. We found that FPSI does not correlate with either of them. This experiment indicates that FPSI is a good measure for functional similarity in promoters.

Genome-Wide Promoter Similarity Profiles Between Orthologs

Using Equation 1, we first examined the frequency distribution of the FPSIs among human genes and their orthologs in other species (Fig. **2**). The number of genes analyzed for each species is available in Supplemental File 1. As can be expected, the highest FPSI occurs between human genes and their orthologs in chimpanzee (Table **1**), with 68% of the orthologs having a FPSI ≥ 0.8. Furthermore, 27% of orthologs have identical promoters (FPSI = 1.0, Supplemental File 2), reminiscent of their phylogenetic closeness. As can be seen in Fig. (**2**), the FPSI profile of the orthologs seen with chimpanzee and other species progressively worsens when humans are compared to other mammals such as mouse and rat, and non-mammals such as chicken and zebrafish. Between human and non-primates, such as mouse, rat and dog, 80% of the orthologs have a FPSI < 0.4 (Fig. **2**, Supplemental File 2). It is even lower between human and non-mammal vertebrates, such as chicken and zebrafish (82% and 93% with a FPSI < 0.3, respectively, Table **1**). This progressive change of overall FPSI among these species are revealed through FPSI between each pair as shown in Table **1**.

The distribution profile of overall FPSI of orthologous genes between chimpanzee and other vertebrates largely resembles that seen between human and these vertebrates (comparison between Figs. **2**, **3A**, Table **1**). It progressively worsens when chimpanzee is compared with other mammals such as mouse and rat, and non-mammals such as chicken and zebrafish.

An interesting finding of this analysis is that the overall FPSI between the two rodents species (i.e., mouse and rat) is quite good at a mean value of 0.374 (underlined in Table **1**), being more than double of each rodent compared to dog (Table **1**, Fig. **3B**). In general, however, the FPSIs between primates and other non-primate mammals (mouse, rat and dog) appear to get progressively worsened as they are compared to non-mammalian vertebrates, such as zebrafish and chicken (Table **1**). Notably, the FPSI is the lowest when the six vertebrates are compared to zebrafish, over 92% below 0.3 (Supplemental File 2).

The closeness of FPSIs among the 20 pairs of comparison is consistent with the *t* statistics (Table **2**). For example,

Table 1. Genome-Wide Mean FPSIs Between Each Pair of Organisms

	Chimpanzee	Mouse	Rat	Dog	Chicken	Zebrafish
Human	0.786	0.218	0.200	0.201	0.130	0.065
Chimpanzee		0.200	0.189	0.192	0.119	0.066
Mouse			0.374	0.167	0.122	0.070
Rat				0.160	0.127	0.074
Dog					0.134	0.073
Chicken						0.057

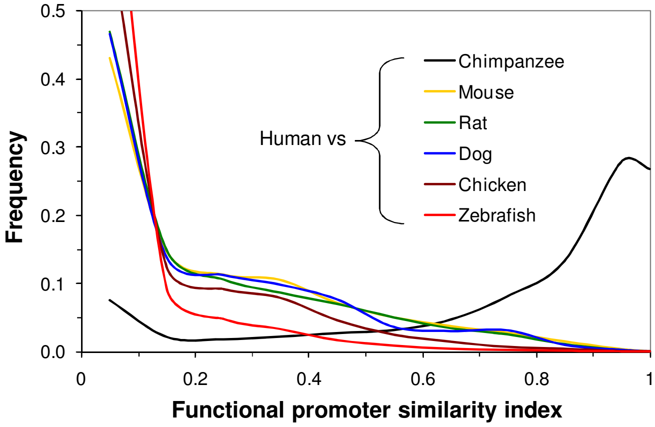

Fig. (2). Genome wide FPSI of orthologous genes between human genes and their orthologs in other vertebrates.

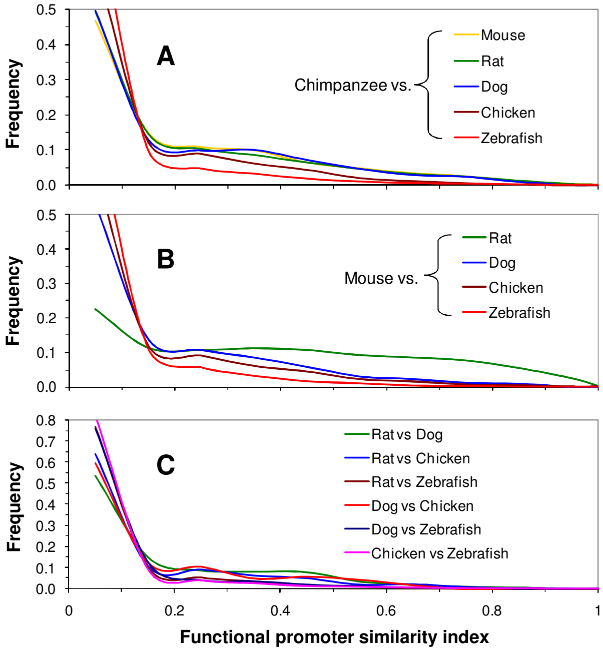

Fig. (3). Genome wide FPSI of orthologous genes in vertebrates other than human.

human and chimpanzee are closest. Their FPSIs are the best and far ahead of the other 19 pairs, the *t* statistics also indicate their FPSIs are drastically different from the other 19 pairs. Similarly, the second closest pair is between the two rodents, mouse and rat. Their FPSIs are second to the human-chimpanzee comparison, and far ahead of the remaining 18 pairs. The *t* statistics indicate that they are significantly different from the other 19 pairs; and their differences from other pairs are not as drastic as those between the two primates compared to the others. On the other hand, when the two primates compared to the three non-primates mammals, their promoter similarities are very close (Rows 1 & 2 in Table **1**). This is also revealed through t statistics (Table **2**). This observation prompted us to study the relationship between the FPSI and time of divergence between the pair of organisms in question. We used the TimeTree Knowledge Base [34, 35] to estimate the divergence time. The promoter similarity appears to be negatively correlated with the time of divergence between the pair (Fig. **4**, Supplemental File **3**).

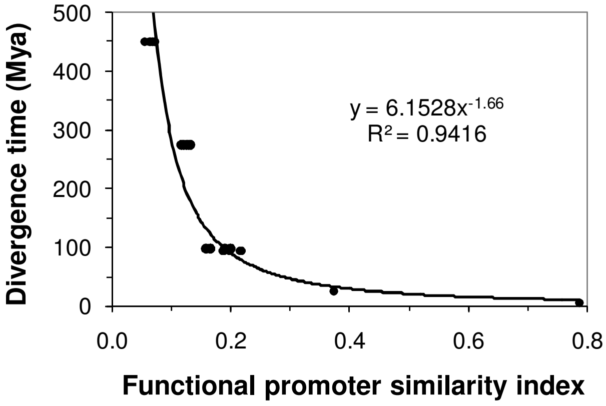

Fig. (4). Correlation between genome-wide FPSI and time of divergence.

Comparison of Human CAGs with their Orthologs in other Vertebrates

The genome-wide FPSI profiles reveal some genes within a genome are more conserved than others. In this regard, we took human CAGs as an example. The comparison between human CAGs (see methods) and their orthologs in chimpanzee indicates that these genes do not show a signifi-

Table 2. *t* Statistics Between FPSIs Among the 20 Pairs of Comparisons

	H&M	H&R	H&D	H&C2	H&Z	C1&M	C1&R	C1&D	C1&C2	C1&Z	M&R	M&D	M&C2	M&Z	R&D	R&C2	R&Z	D&C2	D&Z	C2&Z
H&C1	177.1	133.6	73.1	162.6	241.1	174.8	126.1	68.1	159.4	232.1	81.0	83.8	164.3	238.9	59.3	111.5	186.5	64.2	104.2	202.1
H&M		4.5	2.2	24.7	66.6	6.4	6.7	3.1	26.7	62.2	33.0	7.2	27.0	64.2	5.6	16.3	43.6	8.5	22.1	52.7
H&R			0.1	15.0	35.9	0.1	2.1	1.0	17.1	34.7	31.1	4.3	16.8	34.5	3.8	11.6	28.3	6.4	17.6	33.5
H&D				8.6	17.7	0.1	1.4	0.8	10.0	17.4	19.8	3.3	9.6	17.0	3.2	8.0	15.8	5.4	13.0	18.1
H&C2					19.5	19.0	11.8	6.9	2.6	18.6	45.8	4.8	1.9	18.0	2.7	0.6	13.8	0.3	8.2	18.8
H&Z						54.3	29.8	15.0	15.2	0.5	67.5	14.5	17.0	2.4	9.2	11.3	2.9	6.9	1.3	2.8
C1&M							2.5	1.0	21.2	51.1	36.0	4.7	21.2	52.2	3.9	13.0	36.7	6.7	19.2	44.7
C1&R								0.3	13.8	28.9	31.4	2.8	13.5	28.6	2.7	9.5	24.0	5.3	15.6	28.5
C1&D									8.2	14.8	19.4	2.3	7.8	14.4	2.4	6.6	13.4	4.5	11.4	15.5
C1&C2										14.5	46.9	6.3	0.8	13.8	3.8	1.3	10.5	1.4	6.4	15.1
C1&Z											66.2	14.2	16.2	1.7	9.0	10.9	2.4	6.8	1.1	3.1
M&R												25.3	47.3	66.3	19.3	36.0	58.2	22.4	39.2	63.3
M&D													5.9	13.8	0.6	4.6	12.5	2.8	10.0	15.0
M&C2														15.5	3.5	0.7	11.7	1.1	7.0	16.6
M&Z															8.7	10.4	1.3	6.4	0.5	4.5
R&D																2.9	8.1	1.8	7.2	9.7
R&C2																	8.8	0.6	6.5	11.9
R&Z																		5.9	0.1	4.6
D&C2																			5.2	7.6
D&Z																				2.4

Note: H=human, C1=chimpanzee, M=mouse, R=rat, D=dog, C2=chicken, Z=zebrafish.

Table 3. Mean FPSIs Between Human Cancer-Associated Genes (CAGs), Non-CAGs, and Transcription Factor Genes (TFs), and their Orthologs in Chimpanzee, Mouse, Rat, Dog, Chicken, and Zebrafish

	Chimpanzee	Mouse	Rat	Dog	Chicken	Zebrafish
CAGs	0.782	0.237	0.219	0.236	0.137	0.069
non-CAGs	0.787	0.216	0.197	0.193	0.129	0.064
TFs	0.792	0.330	0.278	0.349	0.225	0.084

Note: CAGs: cancer-associated genes, TFs: transcription factor genes. Underlined: significantly higher (mean ± SE, SE: standard error) than the mean FPSIs of non-CAG or genome-wide mean.

cantly altered FPSI than other genes (Table **3**). However, as we can see from Table **1**, the overall FPSI is already very high between these two primates. Nevertheless, when human CAGs compared with their orthologs in the two rodents and dog, their FPSIs appear to be significantly better than other genes. We then studied the relative distribution of CAGs over all genes in various FPSI regions and found a higher relative density of CAGs at the higher FPSI regions between human CAGs and their orthologs in mouse and rat (Fig. **5**).

The density of CAGs is the proportion of these genes over the entire gene population in the bin. It is noted from Fig. (**5**), panels B and C, that the density appears to increase with an increasing FPSI up to 0.7 and 0.6, respectively, and then it becomes unsteady beyond this point. This unsteadiness is due to the fact that the number of orthologs in these FPSI regions is very small, only about 5% of genes (see Fig. **2**); the values become noisier when the total number of genes in that FPSI region becomes smaller.

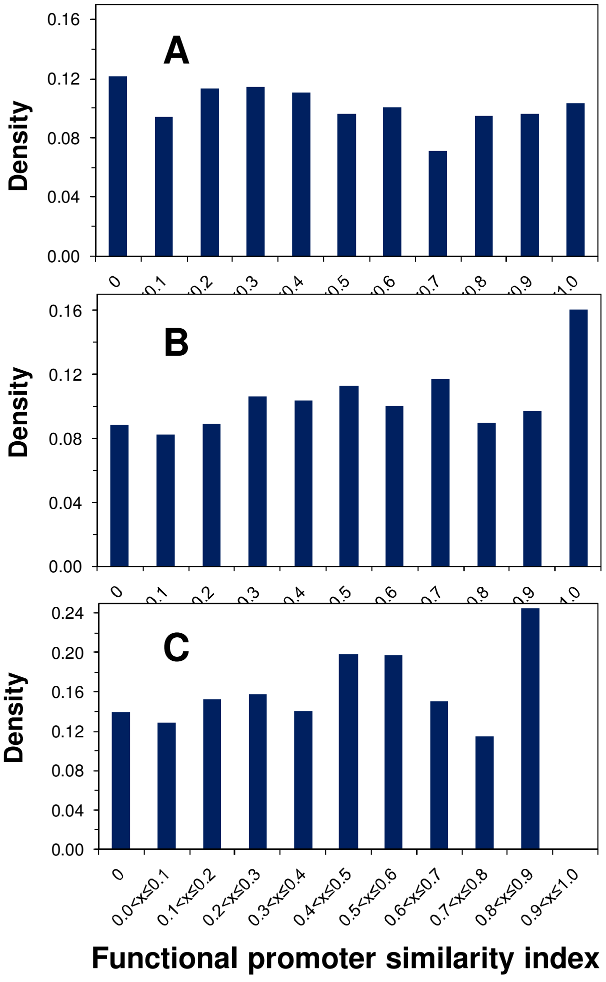

Fig. (5). Density of cancer-associated genes over various regions of the FPSI. **A**: human vs. chimpanzee, **B**: human vs. mouse, **C**: human vs. rat.

The mean values presented in Tables **1** and **3** are the arithmetic average of the individual FPSI values. We also calculated the weighted means and present the result in the Supplemental File 4. The weighted mean values are smaller than those presented in Tables **1** and **3**, but the overall trend is the same.

Functional Characterization

The promoter difference between orthologs appears to be related to the time of divergence between the species. The promoters of CAGs appear to be more conserved than others. In order to find whether the promoter conservation is related with biological functions, we identified two groups of genes, one with more conserved promoters and the other with less conserved promoters. The more conserved gene group consists of 62 human genes whose FPSIs are higher than 0.7 when human is compared with both chimpanzee and mouse. The less conserved gene group consists of 171 human genes whose FPSIs are 0.0 (no common TF) when human is compared with both chimpanzee and mouse (Supplemental File 5). We then performed a gene functional characterization of each group using Gene Ontology AnaLyzer (GOAL) [36]. This reveals that the genes with more conserved promoters are closely related with various developmental process (GO:0007275, GO:0032502, GO:0048856, GO:0048731, GO:0009653, GO:0048513, GO:0001822), while genes with less conserved promoters are related with regulation of transport, protein binding and signalling (GO:0051050, GO:0005515, GO:0007242, GO:0019932, GO:0032501) and their functional representations are less significant even though there are twice number of genes in this group than in the most conserved group (Supplemental File 5).

DISCUSSION

In this study, we applied very stringent threshold ($p \leq 0.05$) in searching the TFBSs based on its probability on a single sequence (1200 bp) [27]. For such reason, we did not consider a motif of 7 bp or less. This study indicates that FPSI is closely related with microarray gene expression modulation among the orthologous genes. But the identity at the sequence level revealed by BLASTN alignment does not correlate with microarray gene expression modulation. This is not surprising because the functional *cis*-regulatory elements imbedded in the promoter are typically short (~6-20 bp) and usually not detected by BLAST. Also, a TF could bind to several very different sequence motifs. For such reason, we chose TFs that bind to these *cis*-regulatory elements to measure the functional similarity of the promoters.

Functional TFBSs are often position specific [29, 37]. Given limited availability of information in known functional regions of some TFBSs, which are not exclusive and our search space are small (1200 bp), we decided not to consider the positions of TFBSs in Equation 1. Instead, we used a very stringent threshold in motif search. Recent research also indicates that some functional TFBSs are phylogenetically conserved across various species and many such methods have been developed over the past decades (see [5] and refs therein). For example, the COmparative Regulatory Genomics (CORG) platform [38] contains promoters and 5' UTR regions of 16,127 groups of orthologous genes from 5 vertebrate species (human, mouse, rat, fugu, and zebrafish). Some TFBSs, such as Erg-1 site, are conserved only within mammals, but diverged in fish. Phylogenetic footprinting in functional motif discovery has been successful. However, it is a fact that many TFs bind to different motifs [5-7]. Given limited information in this regard, restricting motif search by phylogenetic conservation would eliminate some potential association between TFs and their target genes. Such restriction would hinder us from discovering the divergence of promoter function between orthologs. Nevertheless, without considering these two factors described here, our FPSIs are significantly related with microarray gene expression modulation. This indicates that our decision on the measure of functional promoter similarity is sound.

The overall FPSIs between orthologous genes detected in this study are not high. At the outset of this study, we were interested in determining whether gene regulation may account in large part for the phenotypic differences observed amongst species. This study indicates that FPSIs between the orthologs vary significantly among the species studied here. Even in the two closely related primates, human and chimpanzee, only 27% of promoters are identical between the orthologs. This finding that the *cis*-regulatory machineries are less likely conserved among species than their protein coding regions, suggest that regulation of gene products rather than the composition/sequence of these proteins may account for many of the phenotypic differences among species. This result may account for the finding that surveys of developmental gene expression often reveal differences in timing, location and level, even among closely related species (see [39] and refs wherein). Indeed, several cases exist which indicate that some phenotypic changes are more likely to have resulted from *cis*-regulatory mutations than from mutations in the coding regions of a gene (see [16] and refs therein). For example, it was recently reported that only about half of mRNA transcripts of the one-to-one orthologs were detected in placental labyrinth of both human or mouse [40]. We found that the mean FPSI of this group of orthologs is significantly higher than that of the other orthologs. This indicates that the developmental genes are more conserved in promoter, a consistent observation with [41].

Genes encoding TFs are generally highly conserved [42]. In this study, we found that the promoter regions of TF genes are more conserved than other genes (Table **3**). Earlier studies revealed that more than 26% of the known cancer genes are actually TFs [43]. This conservation is evident both in sequence and in function [44]. At the structural level, the DNA-binding domains of many orthologous TFs are very comparable over large phylogenetic distances, allowing them to bind to identical DNA motifs and regulate the same target genes. Additionally, some TFs can bind on different motifs and perform the same function. For example, many *Drosophila* genes with maternally inherited transcripts were found to have alternative promoters utilized later in development [45]; human TF MBD1 can bind on four different motifs (TRASFAC: R25533, R25534, R25535 and R25536) [6]. Our approach in calculating promoter similarity is distinguished by the fact that we choose the calculation to be based on the TFs, which are most likely bound to the promoter, rather than on the DNA motif. This would reduce artefacts caused by the difference at the sequence level due to the fact that the same TF could bind to several *cis*-regulatory elements that are very different at the sequence level [5-7].

However, in the tissue-specific transcriptional regulation, this conservation does not always exist. A recent discovery from liver-specific TFs (FOXA2, HNF1A, HNF4A and HNF6) reveals that the *cis*-regulatory network diverged extensively between mouse and human orthologs. Despite the conserved functions of these TFs, 41-89% of their BSs appear to be species specific [15].

Mutations in the coding region of orthologous genes are well studied in the context of their associations with certain diseases such as cancer [20, 43, 46]. In light of this, we were interested in determining whether promoter regions of the CAGs differed as compared to the majority of other orthologous genes. Our results show that CAGs tended to be more conserved in their promoter functions (Table **3**, Fig. **5**). Previous work shows that cancer genes and essential genes are more conserved in function and in coding sequences than other genes [21-24]. The corresponding conservation of promoter function suggests that regulation of the timing, location and expression levels of these essential genes plays a critical basic role in growth regulation and differentiation across species [39]. This hypothesis is supported by the various roles of them indicating that the CAGs have more conserved promoter functions. These span a range of biological functions (Supplemental File 6) including protein kinase activity (GO:0004672), cell cycle processes (GO:0022402), phosphotransferase activity (GO:0016773), phosphatase activity (GO:0016791) and cell differentiation (GO:0030154). The fact that these genes play different functional roles is in accordance with the view that cancer is caused by defects in genes from multiple functional categories, according to Hanahan and Weinberg's hallmarks of cancer [47].

It is interesting to note that the functional promoter similarity of orthologous genes is correlated with the divergence time of the pair of organisms under consideration. Even though the coding regions of the orthologous genes are very similar or even identical, their promoter regions can be very different. For example, a pair-wise comparison between human and mouse or between human and rat reveals that the FPSIs of the majority (≥70%) of these orthologs are below 0.3 (Fig. **2**, Supplemental File 2). This observation also holds true for the chimpanzee-mouse and chimpanzee-rat comparisons (Fig. **3A**). This prompted us to investigate the promoter similarity between the two rodents: mouse and rat. Not unexpectedly, the FPSI between these two rodent species (mouse and rat) is significantly better than comparing each of them to either of the primates or to dog (Figs. **2**, **3**; Table **1**). This is consistent with their time of divergence as estimated from the TimeTree Knowledge Base [34, 35]. For example, the divergence time between mouse and rat is 25 Mya, which is about 1/4 of that between dog and each of the two rodents (98 Mya). The relative divergence time is inversely proportional to the FPSI, which is more than double (Table **1**, Fig. **4**). It is more interesting to note that the shape of FPSI distribution curve of the Human-Chimpanzee pair is very different from the other pair-wise comparisons (Fig. **2**). This is consistent with the time of divergence. The divergence time between human and chimpanzee is 6.5 Mya, while that of other pair-wise distances (excepting the mouse-rat pair) is at least one order of magnitude higher (Supplemental file 3). The good correlation between FPSI and the divergence time suggests that the FPSI could be a measure for phylogenetic conservation. However, the number of species examined in this study is moderate; this remains to be explored by studying additional organisms and in a broader scale.

This study explored the differences in *cis*- and *trans*-regulatory elements between the orthologous genes. Structural differences, such as chromosomal rearrangements, segmental duplications and copy numbers, are other important contributing factors and worth exploring as indicated in [50-53].

CONCLUSIONS

We proposed a functional promoter similarity index to measure similarity in transcriptional regulation between orthologs in human, chimpanzee, mouse, rat, dog, chicken, and zebrafish. This index is significantly correlated with microarray gene expression modulation. This study shows that the promoter functions are significantly different between orthologous genes although protein coding sequences closely resemble to each other. The CAGs tend to be more conserved in the promoter function as compared with other genes. The high degree of functional promoter similarity of CAGs suggests that their regulation is essential for growth and development across different species.

There is a general understanding that promoter of a gene is less conserved than its coding region [16, 39 and refs therein]. This is evident from studies of individual genes [48, 49]. However, there were not much of genome-wide studies in this regard. This study enlightens that such differences are genome-wide across various vertebrates. We also found that the genome-wide promoter dissimilarity between orthologs is closely correlated with the time of divergence between the organisms under consideration and is higher when comparing human to chimpanzee than when comparing either of the primates to a rodent or another vertebrate. Such close correlation indicates that FPSI could be used as a measure of phylogenetic conservation. This merits further study.

ACKNOWLEDGEMENTS

We are grateful to the two anonymous referees for the constructive comments and criticism on the manuscript. We are thankful to our colleagues Brandon Smith, Institute for Biological Sciences, NRC, and Xuhua Xia, University of Ottawa, for the constructive comments on this work. This research is jointly supported by NRC's Genomics and Health Initiative and NRC's Institute for Information Technology.

ABBREVIATIONS

BS	=	Binding site
CAG	=	Cancer-associated gene
DNA	=	Deoxyribonucleic acid
FPSI	=	Functional promoter similarity index (defined by Equation 1)
NRC	=	National Research Council Canada
PHMM	=	Profile Hidden Markov Model
PWM	=	Positional weight matrix
SNP	=	Single nucleotide polymorphism
TF	=	Transcription factor
TFBS	=	Transcription factor binding site
TSS	=	Transcription start site
UCSC	=	University of California, Santa Cruz
UTR	=	Untranslated region

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] B. Rhead, D. Karolchik, R. M. Kuhn, A. S. Hinrichs, A. S. Zweig, P. A. Fujita, M. Diekhans, K. E. Smith, K. R. Rosenbloom, B. J. Raney, A. Pohl, M. Pheasant, L. R. Meyer, K. Learned, F. Hsu, J. Hillman-Jackson, R. A. Harte, B. Giardine, T. R. Dreszer, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent, "The UCSC Genome Browser Database: update 2010", *Nucleic Acids Res*., vol. 38, pp. D613-D619, 2010.

[2] T. J. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek., "Ensembl 2009". *Nucleic Acids Res*., vol. 37, pp. D690-D697, 2009.

[3] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements", *Nat. Rev. Genet*., vol. 5, pp. 276-287, 2004.

[4] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, "Assessing computational tools for the discovery of transcription factor binding sites", *Nat. Biotechnol*., vol. 23, pp. 137-144, 2005.

[5] Y. Pan, "Advances in the discovery of *cis*-regulatory elements", *Curr. Bioinformatics*, vol. 1, pp. 321-336, 2006.

[6] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender, "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes", *Nucleic Acids Res*., vol. 34, pp. D108-D110, 2006.

[7] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard, “JASPAR: an open-access database for eukaryotic transcription factor binding profiles”, *Nucleic Acids Res*., vol. 32, pp. D91-D94, 2004.

[8] W. M. Fitch, “Distinguishing homologous from analogous proteins”, *Syst. Zool*., vol. 19, pp. 99-113, 1970.

[9] T. Gabaldón, “Large-scale assignment of orthology: back to phylogenetics?” *Genome Biol*., vol. 9, article 235, 2008.

[10] W.M. Fitch, “Homology – a personal view on some of the problems”, *Trend Genet*., vol. 16, pp. 227-231, 2000.

[11] R. A. Jensen, “Orthologs and paralogs – we need to get it right” (Included also respond from J. Gerlt and P. Babbitt), *Genome Biol.*, vol. 2, pinteractions1002.1-1002.3, 2001.

[12] E. V. Koonin, “Orthologs, paralogs, and evolutionary genomics”, *Annu. Rev. Genet*., vol. 39, pp. 309-338, 2005.

[13] E. V. Koonin, “An apology for orthologs – or brave new memes”, *Genome Biol*., vol. 2, comment 1005, 2001.

[14] The International HapMap Consortium, “A second generation human haplotype map of over 3.1 million SNPs”, *Nature*, vol. 449, pp. 851-862, 2007.

[15] D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford, and E. Fraenkel, “Tissue-specific transcriptional regulation has diverged significantly between human and mouse”, *Nat. Genet*., vol. 39, pp. 730-732, 2007.

[16] G. A. Wray “The evolutionary significance of *cis*-regulatory mutations”, *Nat. Genet*., vol. 8, pp. 206-216, 2007.

[17] American Cancer Society, *Cancer facts & figures 2010*. American Cancer Society Inc, Atlanta, 2010.

[18] Canadian Cancer Society’s Steering Committee, *Canadian Cancer Statistics 2010*. Toronto (ISSN 0835-2976), 2010.

[19] B. Vogelstein and K. W. Kinzler, “Cancer genes and the pathways they control”, *Nat. Med*., vol.10, pp. 789-799, 2004.

[20] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, “A consensus of human cancer genes”, *Nat. Rev. Cancer*., vol. 4, pp. 177-183, 2004.

[21] M.A. Thomas, B. Weston, M. Joseph, W. Wu, A. Nekrutenko, and P. J. Tonellato, “Evolutionary dynamics of oncogenes and tumor suppressor genes: higher intensities of purifying selection than other genes”, *Mol. Biol. Evol*., vol. 20, pp. 964-968, 2003.

[22] S.J. Furney, D.G. Higgins, C.A. Ouzounis, and N. López-Bigas, “Structural and functional properties of genes involved in human cancer”, *BMC Genomics*, vol. 7, article 3, 2006.

[23] S. Narsing, Z. Jelsovsky, A. Mbah, and G. Blanck, “Genes that contribute to cancer fusion genes are large and evolutionarily conserved”, *Cancer Genetics and Cytogenetics*, vol. 191, pp. 78-84, 2009.

[24] I.K. Jordan, I.B. Rogozin, Y.I. Wolf, V. Eugene, and E.V. Koonin, “Essential genes are more evolutionarily conserved than are nonessential genes in bacteria”, *Genome Res.*, vol. 12, pp. 962-968, 2002.

[25] C.W. Hay and K. Docherty, “Comparative analysis of insulin gene promoters: implications for diabetes research”, *Diabetes*, vol. 55, pp. 3201-3213, 2006.

[26] S.R. Eddy, “Profile hidden Markov models”, *Bioinformatics*, vol. 14, pp. 755-763, 1998.

[27] Y. Pan and S. Phan, “Threshold for positional weight matrix”, *Eng. Lett.*, vol. 16, pp. 498-504, 2008.

[28] Z. Liu, S. Phan, F. Famili, Y. Pan, A. E. G. Lenferink, C. Cantin, C. Collins, and M.D. O’Connor-Mccourt, “A multi-strategy approach to informative gene identification from gene expression data”, *J. Bioinform. Comput. Biol*., vol. 8, 19-23, 2010.

[29] B. Smith, H. Fang, Y. Pan, P. R. Walker, A. F. Famili, and M. Sikorska, “Evolution of motif variants and positional bias of the cyclic-AMP response element”, *BMC Evol. Biol*., vol. 7, article S15, 2007.

[30] The NCBI Homologene Database. [online] available: ftp://ftp.ncbi.nih.gov/pub/HomoloGene/build63/, [accessed: May 21, 2009].

[31] Q. Cui, Y. Ma, M. Jaramillo, H. Bari, A. Awan, S. Yang, S. Zhang, L. Liu, M. Lu, M. O'Connor-McCourt, E. O. Purisima, and E. Wang, “A map of human cancer signalling”. *Mol. Syst. Biol*., vol. 3, article 152, 2007.

[32] R. S. Stearman, L. Dwyer-Nield, L. Zerbe, S. A. Blaine, Z. Chan, P. A. Bunn Jr, G. L. Johnson, F. R. Hirsch, D. T. Merrick, W. A. Franklin, A. E. Baron, R. L. Keith, R. A. Nemenoff, A. M. Malkinson, and M. W. Geraci, “Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model”, *Am. J. Clin. Pathol*., vol. 167, pp. 1763-1775, 2005.

[33] M. Kapushesky, I. Emam, E. Holloway, P. Kurnosov, A. Zorin, J. Malone, G. Rustici, E. Williams, H. Parkinson, and A. Brazma, “Gene Expression Atlas at the European Bioinformatics Institute”, *Nucleic Acids Res*., vol. 38, pp. D690-D698, 2010.

[34] S.B. Hedges, J. Dudley, and S. Kumar, “TimeTree: a public knowledge-base of divergence times among organisms”, *Bioinformatics*, vol. 22, pp. 2971-2972, 2006.

[35] S.B. Hedges and S. Kumar, Eds., *The Timetree of Life*. Oxford: Oxford University Press, 2009.

[36] A.B. Tchagang, A. Gawronski, H. Bérubé, S. Phan, F. Famili, and Y. Pan, “GOAL: A software tool for assessing biological significance of genes group”, *BMC Bioinformatics*, vol. 11, article 229, 2010.

[37] E. Blanco, X. Messeguer, T. F. Smith, and R. Guigó, “Transcription factor map alignment of promoter regions”, *PLoS Comput. Biol*., vol. 2, article e49, 2006.

[38] C. Dieterich, S. Grossmann, A. Tanzer, S. Röpcke, P. F. Arndt, P. F. Stadler, and M. Vingron, “Comparative promoter region analysis powered by CORG”, *BMC Genomics*, vol. 6, article 24, 2005.

[39] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano, “The evolution of transcriptional regulation in eukaryotes”, *Mol. Biol. Evol*., vol. 20, pp. 1377-1419, 2003.

[40] B. Cox, M. Kotlyar, A. I. Evangelou, V. Ignatchenko, A. Ignatchenko, K. Whiteley, I. Jurisica, S. L. Adamson, J. Rossant, and T. Kislinger, “Comparative systems biology of human and mouse as a tool to guide the modelling of human placental pathology”, *Mol. Syst. Biol*., vol. 5, article 279, 2009.

[41] A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. Edwards, J. E. Cooke, and G. Elgar, “Highly conserved non-coding sequences are associated with vertebrate development”, *PLoS Biol.*, vol. 3, article e7, 2005.

[42] K. S. Zaret, “Regulatory phases of early liver development: paradigms of organogenesis”, *Nat. Rev. Genet*., vol. 3, pp. 499-512, 2002.

[43] X. S. Puente, G. Velasco, A. Gutierrez-Fernandes, J. Berranpetit, M.-C. King, and C. Lopez-Otin, “Comparative analysis of cancer genes in the human and chimpanzee genomes”, *BMC Genomics*, vol. 7, article 15, 2006.

[44] R.P. Zinzen, and E. E. M. Furlong, “Divergence in *cis*-regulatory networks: taking the ‘species’ out of cross-species analysis”, *Genome Biol*., vol. 9, article 240, 2008.

[45] E.A. Rach, H.Y. Yuan, W.H. Majoros, P. Tomancak, and U. Ohler, “Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome”, *Genome Biol*., vol. 10, article R73, 2009.

[46] P.C. Chen, S. Dudley, W. Hagen, D. Dizon, L. Paxton, D. Reichow, S.R. Yoon, K. Yang, N. Arnheim, R. M. Liskay, and S. M. Lipkin, “Contributions by MutL homologues Mlh3 and Pms2 to DNA mismatch repair and tumor suppression in the mouse”, *Cancer Res*., vol. 65, pp. 8662-8670, 2005.

[47] D. Hanahan and R. A. Weinberg, “The hallmarks of cancer”, *Cell*, vol. 100, pp. 57-70, 2000.

[48] V. F. Bumaschny, F. S. de Souza, R. A. López Leal, A. M. Santangelo, M. Baetscher, D. H. Levi, M. J. Low, and M. Rubinstein, “Transcriptional regulation of pituitary POMC is conserved at the vertebrate extremes despite great promoter sequence divergence”, *Mol. Endocrinol*., vol. 21, pp. 2738-2749, 2007.

[49] M. Zhan, T. Miura, X. Xu, and M. S. Rao, “Conservation and variation of gene regulation in embryonic stem cells assessed by comparative genomics”, *Cell Biochem. Biophys.*, vol. 43, pp. 379-405, 2005.

[50] R. Blekhman, A. Oshlack, and Y. Gilad, “Segmental duplications contribute to gene expression differences between humans and chimpanzees”, *Genetics*, vol. 182, pp. 627-630, 2009.

[51] L. Dumas, Y. H. Kim, A. Karimpour-Fard, M. Cox, J. Hopkins, J. R. Pollack, and J. M. Sikela, “Gene copy number variation spanning 60 million years of human and primate evolution”, *Genome Res*., vol. 17, pp.1266-1277, 2007.

[52] L. Hu, D. Segrè, and T. F. Smith, “Evolutionary changes in gene regulation from a comparative analysis of multiple Drosophila species”, *Genome Inform*., vol. 18, pp. 12-21, 2007.

[53] L. Huminiecki and K. H. Wolfe, “Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse”, *Genome Res*., vol. 14, pp. 1870-1879, 2004.

Received: July 13, 2010 Revised: August 24, 2010 Accepted: October 01, 2010