DNA Research **12**, 139–150 (2005)

# Frequency and Distribution of AP-1 Sites in the Human Genome

Huamin Zhou,[1] Tyler Zarubin,[2] Zhiliang Ji,[1] Zheng Min,[1] Wei Zhu,[2] Jocelyn S. Downey,[2] Shengcai Lin,[1] and Jiahuai Han[1,*]

*The Key Laboratory of the Ministry of Education for Cell Biology and Tumor Cell Engineering, School of Life Sciences, Xiamen University, Xiamen 361005, Fujian, China*[1] *and Department of Immunology, The Scripps Research Institute, La Jolla, CA 92037, U.S.A.*[2]

(Received 7 June 2004; revised 26 January 2005)

**Abstract**

The AP-1–binding sequences are promoter/enhancer elements that play an essential role in the induction of many genes in mammalian cells; however, the number of genes containing AP-1 sites remains unknown. In order to better address the overall effect of AP-1 on expression of genes encoded by the entire genome, a genome-wide analysis of the frequency and distribution of AP-1 sites would be useful; yet to date, no such analysis of AP-1 sites or any other promoter/enhancer elements has been performed. We present here our study of the consensus AP-1 site and two single-bp variants showing that the frequency of AP-1 sites in promoter regions is significantly lower than their average rate of occurrence in the whole genomic sequence, as well as the frequency of a random heptanucleotide suggesting that nature has selected for a decrease in the frequency of AP-1 sites in the regulatory regions of genes. In addition, genes containing multiple AP-1 sites are more prevalent than those containing only one copy of an AP-1 site, which again may have evolved to allow for greater signal amplification or integration in the regulation of AP-1 target genes. However, the number of AP-1–regulated genes identified in various studies is far smaller than the number of genes containing potential AP-1 sites, indicating that not all AP-1 sites are activated in a given cell under a given condition, and is consistent with the prediction by others that cellular context determines which AP-1 sites are targeted by AP-1.

**Key words:** AP-1; Promoter/enhancer element; Human genome; Frequency; Distribution

## 1. Introduction

A remarkable fact in any given cell is that regulation of genes in the entire genome can be coordinated with such accuracy that homeostasis can be achieved during cell proliferation, differentiation, and development. Studies have demonstrated that the coordination between gene expression at different levels (transcription, mRNA stabilization, translation, etc.) and the interaction of different regulatory factors at each level enables cells to respond in a purposeful manner to the virtually infinite variety of developmental and environmental signals. One of the most important steps in gene expression is transcription. Transcription is controlled by the combinatorial action of multiple regulatory DNA elements in each gene (promoter/enhancers), and regulatory proteins (transcription factors) that interact with these DNA elements in the regulatory regions. Numerous short DNA sequences (promoter/enhancer elements) that are selectively recognized by different transcription factors have been identified in mammalian genes. The completion of the human genome project has made it possible to determine the frequency and distribution of different promoter/enhancer elements throughout the entire genome and within the promoter regions of different genes. Although an understanding of how cellular regulatory machinery regulates all the genes encoded in the entire genome is a lengthy process that requires studies on all aspects of cellular functions, determination of the frequency and distribution of promoter/enhancer elements is an initial and informative step towards the ultimate understanding of how cells achieve a coordinated genome-wide response in transcription to different environmental and developmental cues.

One of the promoter/enhancer elements that has been extensively studied is the activator protein-1 (AP-1) binding site.[1–3] In fact, the study of the AP-1 site and its regulatory factors played a pioneering role in the history of the elucidation of gene transcriptional regulation.[1,2] AP-1 was first identified as a transcription factor that binds to an essential *cis*-element of the human metallothionein IIa promoter.[1] Soon after, the binding site for

Communicated by Toshihisa Takagi

* To whom correspondence should be addressed. Tel. +1-858-784-8704, Fax. +1-858-784-8665, E-mail: jhan@scripps.edu

AP-1 was also recognized as the TPA (tetradecanoylphorbol-13-acetate) response element (TRE) of several cellular and viral genes.[2] The consensus sequence of the AP-1 binding sites has been defined as TGA(C/G)TCA based on DNAase I protection analyses of TRE elements in diverse genes.[1,2] However, the AP-1–binding sites exhibit some degree of degeneracy.[4] It is now clear that gene transcription via the AP-1–binding site participates in the regulation of a variety of cellular processes including cell proliferation, cell differentiation, cytokine production, apoptosis and oncogenesis.[4–7] The transcription factors known to bind AP-1 sites include *jun* family (c-Jun, JunB, and JunD) and *fos* family members. The *jun* family members can either form homodimers or heterodimers among themselves or dimerize with the *fos* family members. These homodimeric *jun* or heterodimeric *jun-fos* complexes can then bind to AP-1 sites, resulting in enhanced transcription.[4–7] Moreover, it has also been shown that different *jun* or *jun-fos* dimers may have different DNA binding affinities for AP-1 sites and that minor sequence variations of these AP-1 sites may result in selective binding of either *jun* homodimers or *jun-fos* heterodimers.[4,8] AP-1 transcriptional activity is also known to be regulated by the protein levels and post-translational modification of *jun* and *fos* family member proteins.[7] Furthermore, the JNK, Erk and p38 MAP kinase pathways have all been implicated in the regulation of either the protein levels of AP-1 transcription factors or phosphorylation of them, thus yielding a higher order of complexity in the function of the AP-1 regulatory sites.[9–11] Increasing the complexity of AP-1 regulatory site function still further is the growing number of AP-1–interacting proteins that have been identified, which are likely determinants of specific biological responses.[12]

In light of the significance of AP-1 sites, an evaluation of the frequency and distribution of AP-1 sites in the human genome and gene promoter regions should yield useful information for eventually understanding how cellular machinery controls the expression of genes encoded in the entire genome. We first analyzed the frequency and distribution of the consensus AP-1–binding site (TRE) and found that the frequency of TRE in promoter regions is significantly lower than the overall rate of occurrence in the genome as a whole, and the same is also true for the frequency of a random heptanucleotide. Because of the degeneracy in the sequences of AP-1 recognition sites, some variants of the consensus AP-1 site also play a role in AP-1–regulated gene expression.[4,8] We therefore analyzed the frequency and distribution of two variants of the consensus AP-1 site and reached a similar conclusion to that drawn with the analysis of the consensus TRE sequence. Furthermore, we found that genes containing multiple AP-1 sites occur more frequently than would be predicted from an even distribution among AP-1 site-containing genes. Lastly, we showed that the two variants of TRE used in our analysis do, in fact, function as AP-1–binding sites, regulating gene expression in response to extracellular stimuli.

## 2. Materials and Methods

### *2.1. Analysis of AP-1 sites in human genome and promoter regions of genes*

The Blast program in NCBI (bl2seq) was used to identify the frequency of heptanucleotide sequence TGACTCA (TRE), TGAATCA (termed A-TRE) and TGACTAA (termed AA-TRE). In order to identify 100% of the matches for the short heptamers, we set the parameter "expect" value to $10^9$. The blast against the AP-1 sites we selected was performed for sequences of each contig on all 24 human chromosomes. Since there are still gaps in the human genome database, there was a small portion of human genomic sequence that could not be analyzed. The promoters in the genes of randomly picked chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y were analyzed using a 5-kb 5′ sequence directly upstream of the transcription initiation site in each gene and blasted against each of the three AP-1 sequences we selected. A total of 2960 promoters have been analyzed. The pre-masked human genome service (http://repeatmasker.org/cgi-bin/AnnotationRequest) was used to obtain all repeat sequences from chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y. These repeat sequences were then put into the BLAST align program (bl2seq) and blasted against either TRE, A-TRE or AA-TRE, with the "expect" value set to $10^6$ and the "word size" value set to 7.

### *2.2. Statistical Analysis*

Chi-squared tests were performed using the equation $\Sigma(\mathrm{O}-\mathrm{E})^2/\mathrm{E}$ to determine whether observed frequencies (O) differed from expected frequencies (E). Chi-squared values above standard critical Chi-square values for a significant probability level (i.e., 0.05) indicate a deviation from expected frequencies. Correlation analyses were calculated using the equation $\gamma = (\Sigma\mathrm{xy}-(\Sigma\mathrm{x}\Sigma\mathrm{y}/\mathrm{n})\sqrt{(\Sigma\mathrm{x}^2-(\Sigma\mathrm{x})^2/\mathrm{n})(\Sigma\mathrm{y}^2-(\Sigma\mathrm{y})^2/\mathrm{n})}$ where $\gamma$ is the Pearson correlation coefficient and n is the number of chromosome samples. Pearson correlation coefficients indicate the strength of a linear relationship with $-1$ equal to perfect negative correlation and $+1$ equal to perfect positive correlation. The Mann-Whitney Test ($U$) was used to determine whether two populations differ with respect to central tendency using the equation $U_a = \mathrm{n}_a\mathrm{n}_b + \mathrm{n}_a(\mathrm{n}_a+1)/2 - \Sigma\mathrm{R}_a$ where n equals the number of samples from respective group and $\Sigma\mathrm{R}_a$ equals the sum of the ranks from sample a. The populations differ if the smaller of the calculated values of $U$ is equal to or smaller than the table value at the desired level of $\alpha$ (i.e., 0.05 or 0.001). Z-Scores were calculated to determine how far from the mean, in terms of standard deviations, our

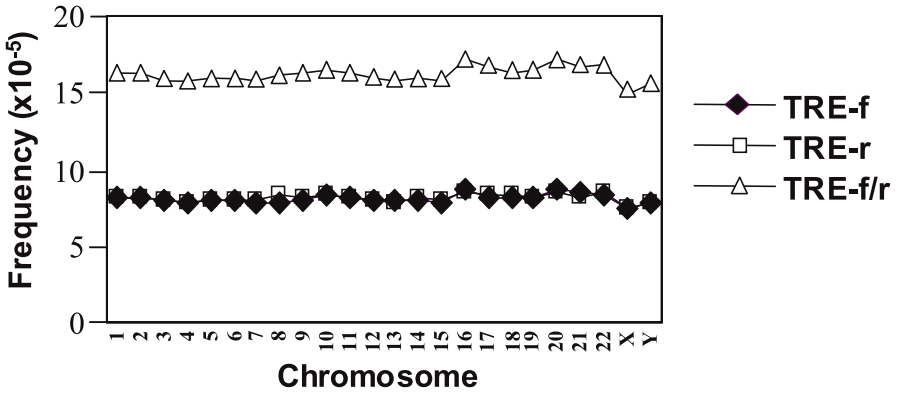


**Figure 1**. Plots of the frequencies of TRE-f, TRE-r, and TRE-f+TRE-r (TRE-f/r) in the 24 different human chromosomes.

observations lie using the equation $Z=(\chi-\mu)/\sigma$, where equals the mean and $\sigma$ equals the standard deviation. With a negative Z-score, the observation lies below the mean and with a positive Z-score, the observation lies above the mean. Confidence intervals (95%) were calculated using the equation $=1.96\sqrt{\{p_x \text{ x } (1\text{-}p_x)/n\}}$[16] where $p_x$ equals the TRE percentage, and n equals the number of analyzed promoter regions.

### *2.3. Reporter vectors*

*Hin*dIII-*Bgl* II oligonucleotides containing seven repeats of TRE, A-TRE, or AA-TRE were inserted into a *cis*-reporter backbone (Stratagene, San Diego, CA).[13]

### *2.4. Cell culture and transfections*

RAW 264.7 cells were maintained in Dulbecco modified Eagle medium plus 10% fetal bovine serum. All plasmid DNA used in transfection experiments was prepared using $CsCl_2$-gradient ultracentrifugation. Any potential LPS and other bacterial sugar/lipid contamination was subsequently removed with Endotoxin Removal Affinity Resin (Associates of Cape Cod, Falmouth, MA). Transfection of RAW 264.7 cells was achieved through the calcium phosphate precipitation technique and included glycerol shock. An empty pcDNA3 vector was used to normalize the amount of total DNA used in each transfection to 3 $\mu$g/well in 6-well plates. LPS (10 ng/ml) or TPA (100 ng/ml) was applied 24 h after transfection for different time periods, as indicated in Fig. 1. The cells were washed in PBS before harvesting, and resuspended in 100 $\mu$l of a reporter lysis buffer (Promega, Madison, WI). Lysed cells were briefly centrifuged, and the relative strength of reporter induction was calculated by measuring the luciferase activity of the supernatant using a luminometer in a luciferase assay reagent (Promega). Transfection efficiency was normalized by co-transfecting cells with an expression plasmid containing a CMV promoter-driven $\beta$-galactosidase reporter. $\beta$-Galactosidase activity was measured by using the chemiluminescent assay Galacto-Light (Tropix, Bedford, MA) or by using *O*-nitrophenyl-$\beta$-D-galactopyranoside (ONPG) as follows: 20 $\mu$l of lysis supernatant was added to 80 $\mu$l of 3.5 mM ONPG solution, incubated at 37°C for 30 min, and absorbance was measured at 405 nm.

## 3. Results

### *3.1. Frequency of consensus AP-1-binding site (TRE) in the human genome*

In order to better assess the contribution of AP-1 transcription factor in overall gene expression, we thought to analyze the frequency and distribution of AP-1 binding sites in the human genome. Since TRE is a double-stranded DNA sequence that reads 5′-TGACTCA-3′ and the other strand 5′-TGAGTCA-3′, we designated them as TRE-f (for forward), and TRE-r (reverse). We analyzed the frequency of TRE-f and TRE-r sequences of TRE independently in each of the 24 human chromosomes. As summarized in Fig. 1, both TRE-f and TRE-r are similarly present in all human chromosomes based on the Chi-squared test ($\chi^2$=0.227 and 0.193, respectively, d.f.=1, $P$ >0.05). Furthermore, there is a strong correlation between the frequency of TRE-f and TRE-r using the Pearson correlation test ($\gamma$ =0.869, n=24, $P$ <0.01), which is consistent with the fact that TRE-f and TRE-r are complementary sequences. The frequencies of TRE-f and TRE-r calculated for the entire human genome were $8.08\times10^{-5}$ and $8.12\times10^{-5}$, respectively. Since the frequency of a random heptanucleotide with the same GC content as TRE is $6.52\times10^{-5}$, based on the published data that genomic GC content is 41%,[14] the frequencies of both TRE-f and TRE-r are substantially higher than the frequency of the random heptanucleotide.

### *3.2. Frequency of TRE in promoter regions*

The available human genome sequence has made it possible to analyze the frequency of TRE in the promoters of different genes. However, it would be very time consuming to analyze the promoters of each of the more than 30,000 genes in the whole genome. Since the

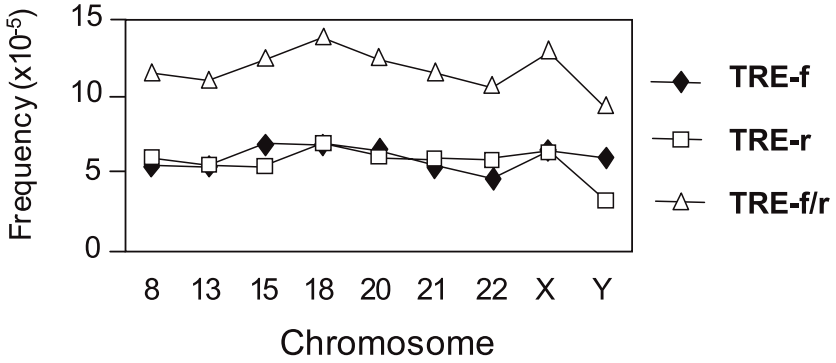


**Figure 2**. The frequencies of TRE-f, TRE-r, and TRE-f+TRE-r (TRE-f/r) in the promoters located on chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y.

genome-wide analysis of TRE sites yielded similar results in the frequency of TRE sites in different chromosomes (Fig. 1), an analysis of several chromosomes for the frequency of TRE sites in promoter regions should provide reliable information regarding the occurrence of TRE in all promoter regions. In order to obtain statistically reliable data we randomly picked chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y for an analysis. We analyzed all defined genes in each of these chromosomes for the frequency of TRE sequences in promoter regions. We selected a 5-kb sequence upstream of the predicted transcriptional initiation sites as the promoter region in our analysis. The frequencies of TRE in the 5-kb promoter region of genes in chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y were calculated and are shown in Fig. 2. The distribution of TRE-f and TRE-r in the promoter regions of genes is similar in the chromosomes analyzed using the Chi-square test ($\chi^2$ =0.728 and 1.431, respectively, d.f.=1, $P$ >0.05). The overall frequencies of TRE-f and TRE-r are $5.95\times10^{-5}$ and $5.93\times10^{-5}$, respectively. These are both 25% less than their frequency in the whole genome ($8.08\times10^{-5}$ and $8.12\times10^{-5}$, respectively). Since GC content in promoters may influence the frequency of TRE in the promoter regions, we measured the average GC content in the promoters analyzed in this study and found it to be 45.8%, which is almost the same as the value of 46% reported in a study of regulatory signals.[14] Therefore, the overall frequencies of TRE-f and TRE-r are also lower than the frequency of a random heptanucleotide ($6.46\times10^{-5}$), based on a GC content of 46% in the promoter region.

We have analyzed statistical significance between the genomic frequency of TRE and the frequency of TRE in the promoter regions (Fig. 3). The $P$-values of Chi square tests were all less than 0.05 when TRE-f (Fig. 3A), TRE-r (Fig. 3B) or TRE-f+TRE-r (Fig. 3C) were analyzed, confirming that TRE has different frequencies in the promoter region compared with the whole genome. The same conclusion was obtained through a Mann-Whitney test. Thus, we concluded that the distribution of TRE is not uniformly distributed in the whole genome and the frequency of TRE in the promoter region is less than the overall frequency of TRE in the whole genome.

### *3.3. Frequency of TRE in repeat-masked sequences*

Sequence repeats are found in varying abundance in most genomes. It is known that these repeats are distributed throughout the genome but their function is largely unknown. We used the pre-masked human genome service in the repeatmasker program (http://repeatmasker.org/cgi-bin/AnnotationRequest) to obtain all repeat sequences from chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y and analyzed the frequency of TRE-f and TRE-r in these repeats. As shown in Fig. 4, the distribution of TRE-f and TRE-r in the repeat-masked regions is similar among the chromosomes when analyzed using a Chi-square test ($\chi^2$ =0.112 and 0.483, respectively, d.f.=1, $P$ >0.05). The overall frequencies of TRE-f and TRE-r are $7.31\times10^{-5}$ and $7.26\times10^{-5}$, respectively. These are higher than their frequency in the promoter regions ($5.95\times10^{-5}$ and $5.93\times10^{-5}$, respectively) and slightly lower than their frequency in the whole genome ($8.08\times10^{-5}$ and $8.12\times10^{-5}$, respectively). There is no statistical significance in the difference between the genomic frequency of TRE and the frequency of TRE in the repeat-masked regions as the $P-$values of Chi square tests are all above 0.05 when TRE-f , TRE-r, or TRE-f+TRE-r were analyzed.

### *3.4. Distribution of TRE in human genes*

We calculated the percentages of genes that contain TRE in their promoter regions and summarized this data in Table 1. Approximately 24% of the analyzed genes contain TRE in the forward direction and 24% in the reverse direction. The calculated frequency of TRE-f (or TRE-r) in the promoters we examined is 25.7% $[1-(1-5.95\times10^{-5})^{5000}]$ if TRE-f (or TRE-r) occurred uniformly in different promoters. Assuming that the analyzed promoter regions (2960 out of $\sim$ 34,000 genes in whole genome) were randomly selected, we can extrapolate with a 95% confidence interval that 24%±1.5% of

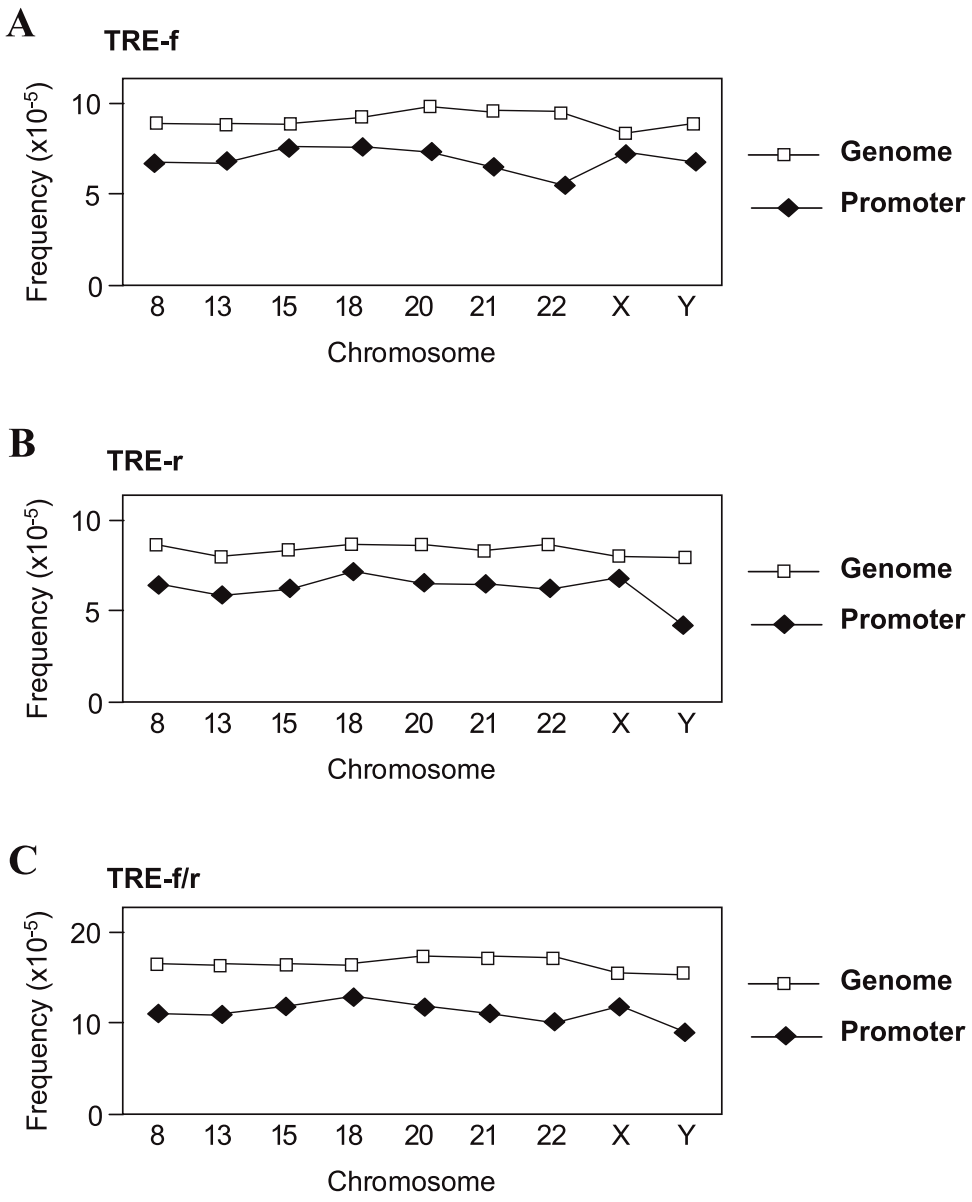


**Figure 3**. Comparison of the frequencies of TRE-f (A), TRE-r (B), and TRE-f+TRE-r (C) in the genome and in promoters.

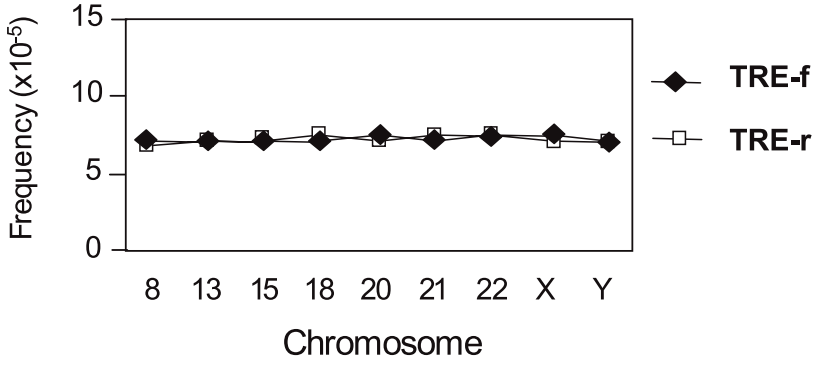


**Figure 4**. The frequencies of TRE-f, TRE-r, and TRE-f+TRE-r in the repeat sequences located on chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y.

**Table 1**. Percentage of genes containing TRE sequence in their promoter.

| Chromosome # | 8 | 13 | 15 | 18 | 20 | 21 | 22 | X | Y | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| **TRE-f** (TGACTCA) | **23** | **24** | **25** | **28** | **26** | **22** | **18** | **25** | **24** | **24** |
| **TRE-r** (TGAGTCA) | **24** | **21** | **24** | **28** | **22** | **29** | **24** | **26** | **13** | **24** |
| **TRE-f** and **TRE-r** | **38** | **35** | **44** | **42** | **39** | **41** | **35** | **40** | **29** | **40** |

**Table 2**. Percentage of promoters containing single and multiple TRE.

| Chromosome # | TRE-f | | | TRE-r | | | TRE-f and/or TRE-r | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1x | 2x | >2x | 1x | 2x | >2x | 1x | 2x | >2x |
| 8 | 18 | 4 | 0.7 | 18 | 5 | 0.5 | 25 | 8 | 5 |
| 13 | 21 | 2 | 1 | 15 | 5 | 0.5 | 21 | 11 | 3 |
| 15 | 18 | 5 | 1 | 19 | 4 | 0 | 30 | 11 | 3 |
| 18 | 23 | 3 | 1 | 23 | 3 | 1 | 25 | 13 | 4 |
| 20 | 19 | 5 | 0.9 | 16 | 4 | 1 | 25 | 10 | 4 |
| 21 | 16 | 5 | 0 | 23 | 5 | 0 | 24 | 15 | 2 |
| 22 | 15 | 3 | 0.3 | 19 | 4 | 0.7 | 23 | 8 | 4 |
| X | 20 | 4 | 0.8 | 21 | 3 | 0.8 | 25 | 11 | 4 |
| Y | 18 | 5 | 0 | 11 | 2 | 0 | 16 | 11 | 2 |
| Overall | 19 | 4 | 0.8 | 18 | 4 | 0.7 | 25 | 11 | 4 |

the gene promoters contained at least one TRE site[15,16] ($p_x$ =0.24, n=2960). The Chi-square test also showed that the difference between 25.7% and 24% is not significant ($P$ >0.05). Thus, the TRE sequences are evenly distributed among the promoters of different genes.

Table 2 summarizes the percentage of genes containing single and multiple TREs in their promoter regions. Using the Chi-square test, all categories except the category of promoters containing one copy of TRE-r ($\chi^2$ =6.83, d.f.=1, $P$ <0.05) show an even distribution among chromosomes ($\chi^2$ =3.84, d.f.=1, $P$ >0.05). Actually, TRE-r is evenly distributed among 8, 13, 15, 18, 20, 21, 22, and X chromosomes and it is unclear whether the low frequency of TRE-r on the Y chromosome has any physiological significance. About 20% of TRE-f — or TRE-r — containing genes have more than one TRE-f or TRE-r, respectively. Since TRE-f and TRE-r are complementary sequences and function identically, the total number of TRE-f and TRE-r is more physiologically relevant. As shown in Table 2, more than one-third of TRE-containing genes have more than one TRE when the direction of TRE was not considered. Indeed, calculating Z-scores associated with binomial distributions for TRE-f and/or TRE-r show that genes with only one copy are the most underrepresented category (Z=−38.8, x=740, $\mu$=1776, $\sigma$=26.65) while those having greater than two copies are the least underrepresented (Z=−7.38, x=118.4, $\mu$=224, $\sigma$=14.4).

### *3.5. Selection of variants of consensus AP-1 recognition sequences for frequency and distribution analysis*

Since the AP-1 transcription factors do not only bind to a consensus DNA sequence but also to some DNA sequences that differ slightly from the consensus AP-1 site,[1,2,4,8] variant AP-1 binding sites need to be considered in order to understanding the overall expression of AP-1-dependent genes. Since TRE possess a twofold rotational (or palindromic) symmetry, a mutation at position 1 or 7, 2 or 6, and 3 or 5 can be considered as the same. Thus, there are ten single-base variants of the TRE sequence; however, not all variants can function as an AP-1 recognition site. For example, substitution of a base at position 1, 3, or 5 has been shown to diminish AP-1 sequence-mediated reporter gene expression.[1,2,17] Since

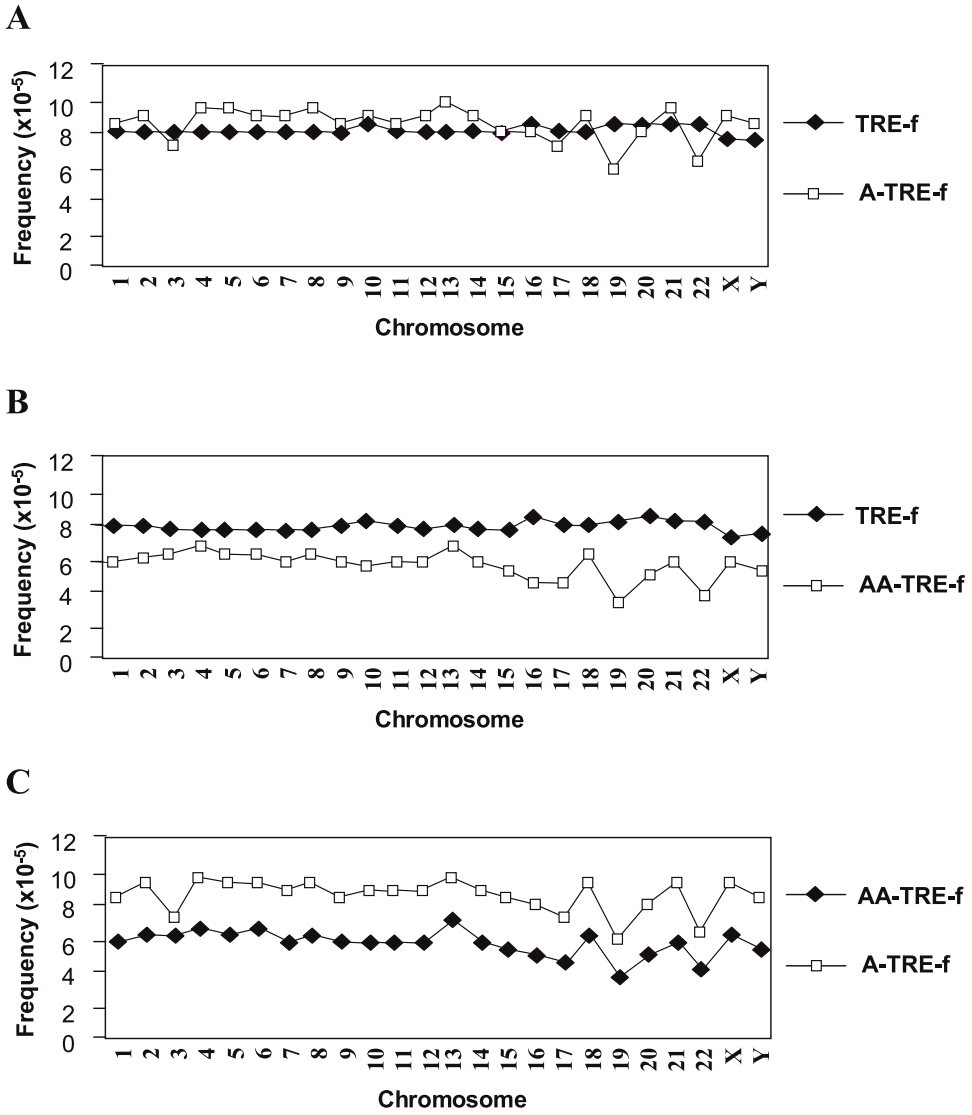


**Figure 5**. Comparison of the genomic frequencies between TRE-f and A-TRE-f (A), TRE-f and AA-TRE-f (B), or AA-TRE-f and A-TRE-f (C).

there is no conclusion as to how many AP-1 site variants can function as AP-1 recognition sites, it is improbable for us to determine the frequency of every AP-1 recognition site. Therefore, we selected for our analysis two AP-1 site variants that are known to function as AP-1 recognition sites. A-TRE, a TRE variant with A instead of C or G at position 4 (TGAATCA) that is reported to be selectively targeted by *jun-fos* heterodimer,[8] was selected in our analysis. Since both TRE and A-TRE are palindromic sequences, we thought to include a non-palindromic AP-1 site variant in our analysis because it may behave differently than the palindromic sequences. We found through a literature search a non-palindromic TRE variant, TGACTAA,[1,18] and included it in our frequency and distribution analysis.

### *3.6. Frequency of A-TRE and AA-TRE in the human genome*

Since the frequencies of TRE-f and TRE-r are about the same (Fig. 1) and our small scale test also showed that the difference of frequency between the two orientations of A-TRE or AA-TRE was less than 0.5% (data not shown), we only analyzed the forward sequences of A-TRE and AA-TRE in the human genome. The overall frequency of A-TRE-f in the human genome is $8.81\times10^{-5}$, lower than the frequency of the random heptanucleotide of the same GC content ($9.38\times10^{-5}$). Pearson correlation analysis indicated that the distribution of A-TRE-f is not correlated with that of TRE-f ($\gamma = -0.369$, n=24, $P < 0.01$) (Fig. 5A). Although the variation of the frequency of A-TRE-f among different chromosomes is larger than that of TRE-f with a range of $5.70\times10^{-5}$ to $9.49\times10^{-5}$, the Chi-square test suggested

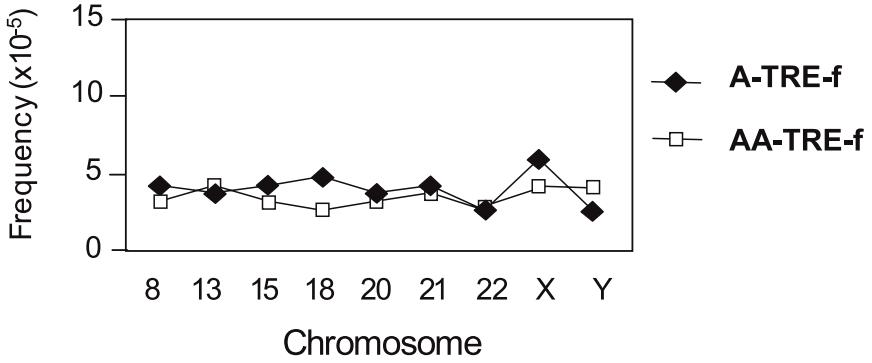


**Figure 6**. The frequencies of A-TRE-f and AA-TRE-f in the promoters located in chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y.

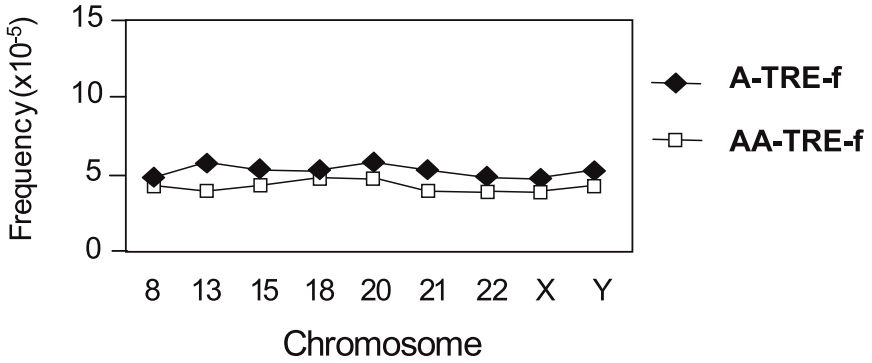


**Figure 7**. The frequencies of A-TRE-f and AA-TRE-f in the repeat sequences located on chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y.

that A-TRE-f still should be considered as uniformly distributed across different chromosomes ($\chi^2$ =2.57, d.f.=1, $P$ >0.05).

The frequency of AA-TRE-f is lower than TRE-f and A-TRE-f in each of the human chromosomes with an overall frequency of $6.07\times10^{-5}$ (Fig. 5B), which is also much lower than the frequency of the random heptanucleotide of the same GC content ($9.38\times10^{-5}$). The distribution of AA-TRE-f in different chromosomes is similar based on the Chi-square test ($\chi^2$ =2.62, d.f.=1, $P$ >0.05). The frequency of AA-TRE-f is not correlated under the Pearson correlation test with that of TRE-f ($\gamma$ = −0.496, n=24, $P$ <0.01) (Fig. 5B), but does show strong correlation with A-TRE-f ($\gamma$ =0.874, n=24, $P$ <0.01) (Fig. 5C).

### *3.7. Frequency of A-TRE and AA-TRE in promoter regions*

We searched for sequences of A-TRE-f and AA-TRE-f in the 5-kb promoter region of genes on chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y. The frequencies of A-TRE-f and AA-TRE-f were calculated and shown in Fig. 6. Both A-TRE-f and AA-TRE-f are evenly distributed in the chromosomes analyzed using a Chi-square test ($\chi^2$ =3.65 and 1.30, respectively, d.f=1, $P$ >0.05). Similar to the frequency and distribution of TRE (Figs. 1 and 2), A-TRE and AA-TRE occurred much less frequently in the promoter regions in comparison with the whole genome (Figs. 5 and 6). The overall frequency of A-TRE-f and AA-TRE-f in the promoter regions is $5.15\times10^{-5}$ and $3.43\times10^{-5}$, respectively. These are lower than the frequency of a random heptanucleotide of the same GC content ($7.59\times10^{-5}$). As mentioned earlier, the lower frequency of AP-1 sites in the promoter regions may be a result of natural selection pressure restricting the number of AP-1 sites in the regulatory region of genes. If this is the case, AA-TRE was most affected in the AP-1 recognition sites we have analyzed.

### *3.8. Frequency of A-TRE and AA-TRE in repeat-masked sequences*

The repeat sequences selected by repeatmasker program from chromosomes 8, 13, 15, 18, 20, 21, 22, X, and Y were analyzed for the frequency of A-TRE-f and AA-TRE-f. As shown in Fig. 7, the distribution of A-TRE-f and AA-TRE-f in the repeat-masked regions is similar in the chromosomes analyzed using a Chi-square test ($\chi^2$ = 0.441 and 0.657, respectively, d.f.=1, $P$ >0.05). The overall frequencies of A-TRE-f and AA-TRE-f are $6.19\times10^{-5}$ and $4.42\times10^{-5}$, respectively. These are higher than their frequency in the promoter regions ($5.15\times10^{-5}$ and $3.43\times10^{-5}$, respectively) and lower

**Table 3**. % of genes containing A-TRE or AA-TRE sequence in their promoter.

| Chromosome # | 8 | 13 | 15 | 18 | 20 | 21 | 22 | X | Y | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| **A-TRE-f** (TGAATCA) | **19** | **19** | **22** | **23** | **21** | **23** | **12** | **29** | **11** | **21** |
| **AA-TRE-f** (TGACTAA) | **14** | **18** | **14** | **11** | **13** | **19** | **9.8** | **18** | **18** | **15** |

**Table 4**. % of promoters containing single or muliple of TRE, A-TRE and/or AA-TRE.

| Chromosome # | 1x | 2x | 3x | >3x | Total |
|---|---|---|---|---|---|
| 8 | 21 | 17 | 6.4 | 5.6 | 50 |
| 13 | 18 | 16 | 10 | 4.8 | 49 |
| 15 | 24 | 18 | 9.7 | 4.2 | 56 |
| 18 | 18 | 21 | 6.9 | 6.3 | 53 |
| 20 | 18 | 20 | 8.5 | 2.9 | 50 |
| 21 | 21 | 22 | 11 | 2.4 | 57 |
| 22 | 19 | 14 | 6.2 | 2.2 | 42 |
| X | 19 | 20 | 12 | 4.6 | 56 |
| Y | 29 | 13 | 7.9 | 0 | 50 |
| Overall | 19 | 18 | 10 | 5.0 | 52 |

than their frequency in the whole genome ($8.81\times10^{-5}$ and $6.07\times10^{-5}$, respectively). The differences are all statistically significant as the $P$-values of Chi-square tests are all less than 0.05.

### *3.9. Distribution of A-TRE and AA-TRE in human genes*

We calculated the percentages of genes that contain A-TRE or AA-TRE in their promoter regions and summarized this data in Table 3. The percentage of genes containing A-TRE or AA-TRE is 21% and 15%, respectively. These frequencies are about the same as the 22% $[1-(1-5.15\times10^{-5})^{5000}]$ and 15% $[1-(1-3.43\times10^{-5})^{5000}]$ that a uniform distribution would predict.

Because the consensus AP-1 site and other AP-1 recognition sites may function together in a given promoter, we analyzed the percentage of the promoters containing one, two, three, and more than three TRE, A-TRE-f, and/or AA-TRE-f (Table 4). As shown in Table 4, the number of promoters containing one of the three TREs is only slightly higher than the promoters containing two. Furthermore, the percentage of promoters that contain two or more TREs is greater than the percentage of promoters containing a single TRE element. It should be noted that the percentage of promoters containing single ($1\times$) TRE, A-TRE and/or AA-TRE (Table 4) is even smaller than the percentage of promoters containing $1\times$ TRE (Table 2) on some chromosomes, which is due to the fact that a promoter containing one type of TRE may also contain other TRE variants. The data in Table 4 is incomplete because we cannot include all potential AP-1 sites in our analysis, but it still can be concluded that more than half of all genes contain potential AP-1 recognition sites. Each of the TRE sequences we analyzed appears to be uniformly distributed in different promoters (Tables 1 and 3), however, the distribution of different AP-1 sites seems not to be independent as they tend to be concentrated on some genes (Table 4). We have compared all categories in Table 4 with predicted binomial distributions. Z-Scores for both the $1\times(-437)$ and $2\times$ categories $(-11.75)$ are low, suggesting that single or double copies of AP-1 sites are underrepresented in the promoter region. In contrast, Z-scores for the $3\times(+4.89)$ and $>3\times(+6.32)$ imply an overrepresentation in promoter regions containing three or more AP-1 sites. Different numbers of AP-1 recognition sites in different genes, then, may be a mechanism that allows for the fine modulation of gene expression by AP-1 transcription factors.

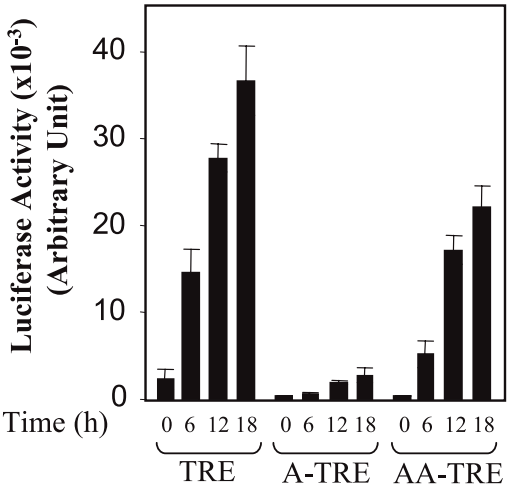


**Figure 8**. Comparison of TRE-, A-TRE- and AA-TRE-dependent gene expression. RAW264.7 cells were transfected with the reporter plasmids of TRE, A-TRE, or AA-TRE. Twenty-four hours after transfection, the cells were treated with LPS (10 ng/ml) for different periods of time as indicated. The cells were harvested and the luciferase activity was measured.

### *3.10. Comparison of TRE-, A-TRE- and AA-TRE-dependent gene expression*

Since we had analyzed frequencies and distributions of TRE, A-TRE and AA-TRE, we believed a functional comparison of these AP-1 sites would provide important information on the relationship between the different sequences and their regulatory behaviors. To compare the function of these three AP-1 sites in gene expression, we constructed expression vectors of a luciferase reporter under the control of TRE, A-TRE, or AA-TRE, respectively. Since it is known that AP-1 transcriptional activity can be induced in macrophages by LPS,[19] we transiently transfected RAW 264.7 cells with TRE, A-TRE, or AA-TRE reporter plasmids. The reporter gene expression was measured in cells treated with or without LPS (Fig. 8). The basal expression of TRE-dependent genes was over fivefold higher than that of A-TRE– and AA-TRE–dependent genes. Although LPS stimulation increased gene expression in a time-dependent manner for each of these three reporters, the levels of induction were significantly different. The TRE reporter construct exhibited the highest luciferase expression levels after LPS stimulation. AA-TRE–mediated luciferase expression was about 60% of TRE expression while A-TRE–mediated luciferase expression was dramatically lower than that directed by TRE or AA-TRE. Due to high basal expression levels, the fold of induction by LPS of the TRE reporter was only one-third of that of the AA-TRE reporter. The differences among TRE, A-TRE, and AA-TRE in mediating reporter gene expression were similarly seen in the cells treated with TPA (data not shown). Thus, the nature of different AP-1 recognition sites as well as their distribution are important factors in the accurate regulation of genes in biological processes.

## 4. Discussion

By virtue of their oncogenic origins,[20] the AP-1 factors Jun and Fos play crucial roles in controlling homeostasis of cell growth.[7] They must receive numerous signals to either upregulate or downregulate gene transcription, depending on cellular context and signals. Our analysis of the human genome database revealed that the occurrence of AP-1 binding sites in promoter regions of different genes is less frequent compared with the overall rate of occurrence in the whole genome (Figs. 1–3, 5, and 6), suggesting that natural selection has eliminated AP-1 binding sites in the promoter region of certain genes. The lower frequency of AP-1 sites in repeat sequences in comparison with the overall rate in the whole genome (Figs. 4 and 7) may be interpreted by the suggestion that interspersed repeats are one of the hotspots for eukaryotic chromosome evolution.[21] Perhaps because the repeats are only loosely related with gene expression,[22] the frequency of AP-1 sites in the repeats is still higher than that in the promoter regions (Figs. 3, 4, 6, and 7). Although the frequency of AP-1 binding sites in the promoter regions is less than expected, there is still a considerable number of genes that have potential AP-1 recognition sites in their promoter regions based on the observation that about one-third of genes contain the consensus AP-1 site in their promoter regions (Table 2). More than half of the genes contain AP-1 recognition sites if variants of the consensus AP-1 site are included (Table 4). In addition, two-thirds of the genes that contain potential recognition sites have more than one site (Table 4). Furthermore, our data indicate that the occurrence of AP-1 sites is not evenly distributed in the human genome and it is likely that these sites have undergone natural selection in order to meet the demand of controlled gene expression in the cell (Table 4). Since the number and diversity of

genes containing AP-1 sites is so large, it is impossible to precisely categorize what types of gene will contain AP-1 sites. Nevertheless, the information regarding the frequency and distribution of AP-1 sites obtained in our analysis has provided some useful information for understanding the overall contribution of AP-1 sites in gene regulation.

A number of AP-1 site variants have been reported to function as AP-1 binding sites and some differ from the consensus TRE sequence by two or more bases. We have limited our analysis to single-base variants of TRE because sites of further deviation behave drastically different from the consensus AP-1 sequence[4] and are beyond the scope of the present study. Although calculating the percentage of genes containing multiple sites of TRE, A-TRE, or AA-TRE cannot give a precise answer as to the percentage of genes containing multiple copies of all AP-1 recognition sites, this analysis does, however, provide solid evidence that at least half of the human genes contain potential AP-1 binding sites in their 5-kb upstream sequence of their transcriptional initiation site. In addition, our data are reliable for interpreting whether AP-1 sites tend to be uniformly distributed or selectively concentrated in particular genes.

It is known that many types of genes, such as oncogenes and genes coding for cytokines and metalloproteinases, are targeted by AP-1 transcription factors; however, it is still not clear whether AP-1 recognition sites are concentrated on certain type of genes compared with the rest of the genome. AP-1 sites are indeed found in cytokine genes, oncogenes, and metalloproteinase genes; however, there are many other types of genes — such as defensins, translation factors, ring-finger proteins — that contain AP-1 sites as well in their promoters. Perhaps as a result of gene duplication, AP-1 sites are often found in all or most members of a gene family. For example, promoters of the 26 solute carrier family members located on chromosomes 8, 13, 15, 18, 20, 22, and X all contain potential AP-1 recognition sites. Thus, AP-1 sites are more frequently found in certain genes.

The results of various experiments indicate that the number of genes induced by AP-1 factors in any given condition is far less than half the number of genes in the cell.[4,23–25] Thus, only a small subset of genes that contain potential AP-1 regulatory sites are activated by AP-1 transcription factors in a particular cell at a given time. This is not surprising since many AP-1 sites may not be accessible due to chromatin structures. Furthermore, gene expression in mammalian cells is controlled by multiple transcription factors, and the binding of the AP-1 transcription factor to a gene may not be sufficient by itself to turn on transcription. Since the sequences surrounding AP-1 sites have been shown to play an important role in directing the orientation of AP-1 transcription factor binding[26] and the interaction of AP-1 transcription factor with other transcription factors,[4] the flanking sequence of an AP-1 site(s) in a given gene needs to be considered. It is clear that whether or not a potential AP-1 site functions in gene expression should depend on the particular cellular context, while the identification of AP-1 sites provides a basis for further analysis.

## References


1. Lee, W., Mitchell, P., and Tjian, R. 1987, Purified transcription factor AP-1 interacts with TPA-inducible enhancer elements, *Cell*, **49**, 741.
2. Angel, P., Imagawa, M., Chiu, R. et al. 1987, Phorbol ester-inducible genes contain a common cis element recognized by a TPA-modulated trans-acting factor, *Cell*, **49**, 729.
3. Angel, P. and Karin, M. 1991, The role of Jun, Fos and the AP-1 complex in cell-proliferation and transformation, *Biochim. Biophys. Acta*, **1072**, 129.
4. Chinenov, Y. and Kerppola, T. K. 2001, Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity, *Oncogene*, **20**, 2438.
5. Toone, W. M., Morgan, B. A., and Jones, N. 2001, Redox control of AP-1-like factors in yeast and beyond, *Oncogene*, **20**, 2336.
6. Mechta-Grigoriou, F., Gerald, D., and Yaniv, M. 2001, The mammalian Jun proteins: redundancy and specificity, *Oncogene*, **20**, 2378.
7. Shaulian, E. and Karin, M. 2001, AP-1 in cell proliferation and survival, *Oncogene* **20**, 2390.
8. McBride, K. and Nemer, M. 1998, The C-terminal domain of c-fos is required for activation of an AP-1 site specific for jun-fos heterodimers, *Mol. Cell Biol.*, **18**, 5073.
9. Whitmarsh, A. J. and Davis, R. J. 1996, Transcription factor AP-1 regulation by mitogen-activated protein kinase signal transduction pathways, *J. Mol. Med.*, **74**, 589.
10. Pramanik, R., Qi, X., Borowicz, S. et al. 2003, p38 isoforms have opposite effects on AP-1-dependent transcription through regulation of c-Jun. The determinant roles of the isoforms in the p38 MAPK signal specificity, *J. Biol. Chem.*, **278**, 4831.
11. Chen, R. H., Abate, C., and Blenis, J. 1993, Phosphorylation of the c-Fos transrepression domain by mitogen-activated protein kinase and 90-kDa ribosomal S6 kinase, *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 10952.
12. van Dam, H. and Castellazzi, M. 2001, Distinct roles of Jun : Fos and Jun : ATF dimers in oncogenesis, *Oncogene*, **20**, 2453.
13. Zhu, W., Downey, J. S., Gu, J., di Padova, F., Gram, H., and Han, J. 2000, Regulation of TNF Expression by Multiple Mitogen-Activated Protein Kinase Pathways, *J. Immunol.*, **164**, 6349.
14. Kel-Margoulis, O. V., Tchekmenev, D., Kel, A. E. et al. 2003, Composition-sensitive analysis of the human genome for regulatory signals, *In Silico. Biol.*, **3**, 145.
15. Harding, D. 1992, Political Polls and Errors, *Teaching Statistics*, **14**, 6.
16. Kmietowicz, Z. W. 1990, The significance of the Lead in Opinion Polls, *Journal of Applied Statistics*, **17**, 9.
17. Kawasaki, N., Satonaka, M., Imagawa, M., Naito, H., and Kawasaki, T. 1998, Functional characterization of the

bovine conglutinin promoter: presence of a novel element for transcriptional regulation of a C-type mammalian lectin containing a collagen-like domain, *J. Biochem.*, (Tokyo) **124**, 1188.
18. Moulton, K. S., Semple, K., Wu, H., and Glass, C. K. 1994, Cell-specific expression of the macrophage scavenger receptor gene is dependent on PU.1 and a composite AP-1/ets motif, *Mol. Cell Biol.*, **14**, 4408.
19. Mackman, N., Brand, K., and Edgington, T. S. 1991, Lipopolysaccharide-mediated transcriptional activation of the human tissue factor gene in THP-1 monocytic cells requires both activator protein 1 and nuclear factor kappa B binding sites, *J. Exp. Med.*, **174**, 1517.
20. Vogt, P. K. 2001, Jun, the oncoprotein, *Oncogene*, **20**, 2365.
21. Eichler, E. E. and Sankoff, D. 2003, Structural dynamics of eukaryotic chromosome evolution, *Science*, **301**, 793.
22. Oei, S. L., Babich, V. S., Kazakov, V. I., Usmanova, N. M., Kropotov, A. V., and Tomilin, N. V. 2004, Clusters of regulatory signals for RNA polymerase II transcription associated with Alu family repeats and CpG islands in human promoters, *Genomics*, **83**, 873.
23. Fu, S. L., Waha, A., and Vogt, P. K. 2000, Identification and characterization of genes upregulated in cells transformed by v-Jun, *Oncogene*, **19**, 3537.
24. Jain, J., Nalefski, E. A., McCaffrey, P. G. et al. 1994, Normal peripheral T-cell function in c-Fos-deficient mice, *Mol. Cell Biol.*, **14**, 1566.
25. Chen, J., Stewart, V., Spyrou, G., Hilberg, F., Wagner, E. F., and Alt, F. W. 1994, Generation of normal T and B lymphocytes by c-jun deficient embryonic stem cells, *Immunity*, **1**, 65.
26. Rajaram, N. and Kerppola, T. K. 1997, DNA bending by Fos-Jun and the orientation of heterodimer binding depend on the sequence of the AP-1 site, *EMBO J.*, **16**, 2917.