Genomics & Informatics Vol. 1(1) 40-49, September 2003

Computational Analysis of Neighboring Genes on *Arabidopsis thaliana* Chromosomes 4 and 5: Their Genomic Association as Functional Subunits

Sung-Ho Goh, Tae-Hyung Kim, Jee-Hyub Kim, DouGu Nam, Doil Choi, and Cheol-Goo Hur*

Laboratory of Plant Genomics Center, Korea Research Institute of Bioscience and Biotechnology, Taejon, Korea

Abstract

The genes related to specific events or pathways in bacteria are frequently localized proximate to the genome of their neighbors, as with the structures known as operon, but eukaryotic genes seem to be independent of their neighbors, and are dispersed randomly throughout genomes. Although cases are rare, the findings from structures similar to prokaryotic operons in the nematode genome, and the clustering of housekeeping genes on human genome, lead us to assess the genomic association of genes as functional subunits. We evaluated the genomic association of neighboring genes on chromosomes 4 and 5 of Arabidopsis thaliana with and without respectively consideration of the scaffold/matrix-attached regions (S/MAR) loci. The observed number of functionally identical bigrams and trigrams were significantly higher than expected, and these results were verified statistically by calculating *p*-values for weighted random distributions. The observed frequency of functionally identical bigrams and trigrams were much higher in chromosome 4 than in chromosome 5, but the frequencies with, and without, consideration of the S/MAR in each chromosome were similar. In this study, a genomic association among functionally related neighboring genes in *Arabidopsis thaliana* was suggested.

Keywords: Scaffold/matrix-attached regions# (S/MAR), bigram, trigram, weighted random distribution

*Corresponding author: E-mail hurlee@kribb.re.kr
Tel +82-42- 879-8560, FAX +82-42-879-8569
Abbreviations: bigram, cluster of two genes, trigram, cluster of three genes.
Accepted 2 May 2003

Introduction

The more genomes of various organisms are revealed, from complete sequencing, the more insights we gain into the sequences themselves. These include: the organization of genomes, the structure of genes and regulatory elements, and the conservation of gene order in evolution (Dandekar *et al.*, 1998). The genome rearrangement in living organism is a progressive form of evolution, where genomes are constantly rearranged and shuffled (Von Mering and Bork, 2002). In bacterial genomes, the strength of genomic associations correlates with the strength of the functional associations between the genes. Several reports have suggested that genomic associations reflect functional association between their proteins (Dandekar, 1998; Enright *et al.*, 1999; Marcotte *et al.*, 1999; Pellegrini *et al.*, 1999; Overbeek *et al.*, 1999; Huynen *et al.*, 2000; Yanai *et al.*, 2001). In addition, Snel *et al.* (2002) obtained a protein interaction network by combining the pairwise interactions between proteins, predicted from the conserved co-occurrence of their genes in operons (Snel *et al.*, 2002). The genomes of higher-order eukaryotes, like animals, plants and fungi, seem to be relatively disorganized, with the average gene generally assumed to be independent of its neighbors, with only a few exceptions, such as repeats of similar sequences caused by gene duplications, and a limited number of ancient gene clusters containing functionally related genes (Von Mering and Bork, 2002). However, it has been revealed that neighboring genes are occasionally assembled into regulatory units, called operons, in the nematode (Blumental *et al.*, 2002). The estimated proportion of genes, expressed as a part of operon, in *Caenorhabditis elegans* was 13-15% (Blumental *et al.*, 2002). In addition, correlation between transcriptome and protein-protein interactions was mapped for *Saccharomyces cerevisiae*, with genes from the same functional cluster showing a higher protein interaction density (Ge *et al.*, 2001). In this respect, it is plausible that genes with similar transcription profiles may have a tendency to cluster in eukaryotic genomes (Cohen *et al.*, 2002; Lercher *et al.*, 2002), and it is suggested that functionally related proteins, encoded by neighboring genes, either physically interact or are involved in a certain biological event. Although eukaryotic genes are not exactly the same as bacterial operons, it would be advantageous

for the sets of genes involved in a certain biological process, to be localized as neighbors on the genome, with some conservation of gene order (Lercher *et al.*, 2002), where their expression might be regulated as a functional module.

Eukaryotic chromosomes at the interphase do not exist as condensed structures, but their relaxed chromatin is attached on the scaffold/matrix of the nucleus, and the looped structure can be dealt with as a functional domain of the chromosome or genome (Liebich *et al.*, 2002). Efforts to reveal the relationship between gene regulatory mechanisms and the nuclear architecture have proved increased evidence (Stein, 1998), and the scaffold/matrix-attached region (S/MAR) has been suggested as one of the abundant regulatory DNA elements of the eukaryotic genome (Frish *et al.*, 2001). S/MAR form the anchor points of loop domains, with domain sizes ranging from a few kilo bases, to more than one hundred (Bode *et al.*, 1992), harbor one or more genes. However, there is no information on either the average gene number, or the functional relatedness between neighboring genes in a loop.

S/MARt DB deposits several hundred S/MAR containing sequences, extracted from original publications (Liebich *et al.*, 2002), and several bioinformatics methods form in *silico* S/MAR prediction have been developed programs, such as SMARTest (Frish *et al.*, 2001) and MAR-Finder (Singh *et al.*, 1997). These tools use several motifs in their library, including origin of replication, TG-rich sequences, curved DNA, linked DNA, topoisomerase II sites, and AT-rich sequences. These motifs, however, do not always appear on every known S/MAR containing sequences. Previously reported S/MAR consensus patterns were recently compared for their enrichment, and their MAR/SAR recognition signature (MRS) (Van Drunen *et al.*, 1997; Van Drunen *et al.*, 1999) verified as the most enriched *motifs* in the S/MAR containing sequences (Liebich *et al.*, 2002).

In the present study, we collected neighboring gene sets, with and without considering the S/MAR from chromosomes 4 and 5 of *Arabidopsis thaliana*, then analyzed the relation between their genomic and functional associations. The effects of S/MAR on the association of genes, with identical function sub-categories, were not confirmed, but it was suggested that genes in the same functional sub-category were assembled together, and with statistical significance.

Results

Bigrams without considering S/MAR

In chromosomes 4 and 5 of *Arabidopsis thalinana*, there are 3744 (Mayer *et al*,. 1999) and 5874 (The Kazusa DNA Research Institute, 2002) non-overlapping genes, respectively. We collected bigrams from two frames according to the starting position. According to the MAtDB, however, some genes had more than one *function category* code assigned. These might come either from the ambiguity of the gene function making annotators difficult to define exactly, or from the multifunctional feature of gene products. Therefore, there is some increase in the number of bigrams and trigrams caused by combination between redundant functions, but these additional function assignments should not be ignored. We collected and accepted all the additional combined bigrams. The number of bigrams was 2178 for each frame in chromosome 4, and 3208 and 3213 for frame1 and 2 of chromosome 5 (Table 1). The average proportion of bigrams, excluding the 00, 98, and 99 function categories were 17.21 and 10.87% for chromosome 4 and 5, respectively (Table 1). From these, functionally identical bigrams, that is, bigrams composed of an identical function sub-category, were counted, and the proportions were 4.59 and 4.68% for frames 1 and 2 of chromosome 4, and 2.46 and 2.33% for frames 1 and 2 of chromosome 5, respectively (Table 1).

We evaluated the *p*-values of functionally identical bigrams to assess the statistical significance for a weighted random distribution. The *p*-values for frames 1 and 2 for chromosome 4 were 1.8279×10^{-95} and 1.219293×10^{-93}, and for chromosome 5 were 1.7615×10^{-47} and 1.3772×10^{-41}, respectively (Table 1). These *p*-values suggested that the probability of a genomic association of functionally identical bigrams, due to chance, was extremely low. The observed number of functionally identical bigrams was significantly higher than expected, even when the weighted random distribution was considered (Table 1). The observed number was higher in chromosome 4, although the chromosome 5 also had a higher than expected number.

We mapped all the bigrams on the diagonal matrices according to their nineteen large function categories in order to display their global genomic association (Fig. 2). The diagonal pairs showed higher frequencies of bigrams, and the pairs on categories '01-metabolism' and '04-transcription' showed relatively higher frequencies than the other pairs. The metabolism and transcription categories are the first and second largest groups in both chromosomes 4 (01: 9.6%, 04: 5.6%) (Mayer *et al.*, 1999) 5 (01: 21.1%, 04: 18.6%) (The Kazusa DNA Research Institute, 2002). Thus, it is plausible that those associated pairs would appear more frequently. However, the diagonal pairs, i.e., composed of the same function category, appeared more frequently regardless of their function category and the proportion of the function category in each chromosome. This coincided with higher probability of co-localizations of genes, composed of identical function sub-categories, as functionally identical bigrams. The

Table 1. Statistics on bigrams of chromosomes 4 and 5 without considering S/MAR.

	Chromosome 4			Chromosome 5		
	Frame 1	Frame 2	Total	Frame 1	Frame 2	Total
Expected No. of functionally identical bigrams	16.7720	17.5968		18.3054	18.3054	
Observed No. of functionally identical bigrams	100 (4.59%)	102 (4.68%)	202 (4.64%)	79 (2.46%)	75 (2.33%)	154 (2.40%)
No. of bigrams w/o 00/98/99 categories	366 (16.80%)	384 (17.63%)	750 (17.21%)	349 (10.88%)	349 (10.86%)	698 (10.87%)
No of total Bigrams	2178	2178	4356	3208	3213	6421
P-value	P(Z>20.8048) $\approx 1.8279 \times 10^{-95}$	P(Z>20.5982) $\approx 1.3293 \times 10^{-93}$		P(Z>14.5733) $\approx 1.7615 \times 10^{-47}$	P(Z>13.6129) $\approx 1.3772 \times 10^{-41}$	

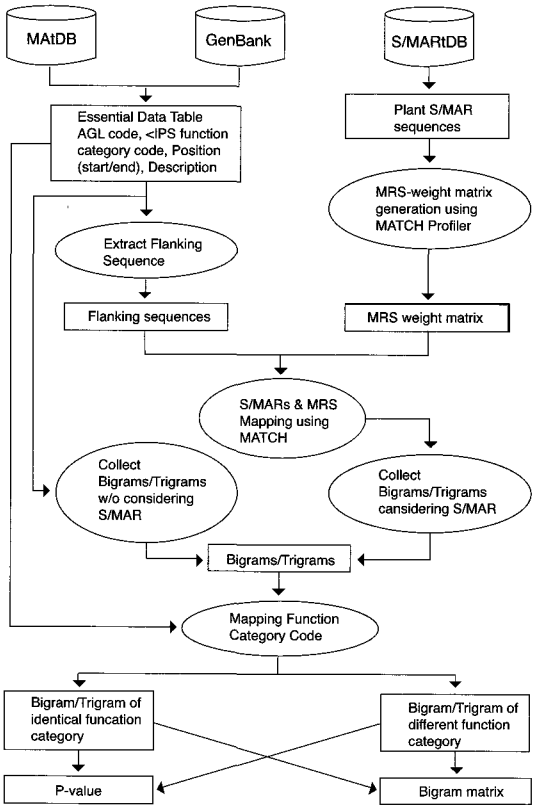

Fig. 1. Analysis flow of neighboring genes.

matrices of chromosome 4 showed clearer, denser pairs on the diagonal than those of chromosome 5, and the pairs related to the metabolism-01 and transcription-04 function categories showed a higher frequency than the other pairs. This occurred because there were more fully annotated genes in chromosome 4 than in chromosome 5, and there was bias in the proportion of function categories of annotated genes.

Prediction flanking sequences that containing S/MAR locus and collection of bigrams

To assess the effect of S/MAR on the co-localization of genes with an identical function sub-category, we predicted the S/MAR loci on chromosomes 4 and 5, and surveyed the bigrams on both sides of S/MAR. We collected the flanking sequences of he discrete non-overlapping genes, then assessed their S/MAR retention. The MATCH™ Profiler program generated five criteria for MRS-1 and MRS-2, and the cutoff value FN50 was selected following tests on the previously reported sequences. The sequences used for these tests were the plastocyanin (z83321), ATB2 (z82043) and *ATH1* (z83320) genes of *Arabidopsis thaliana*, and they experimentally confirmed for their S/MAR retention (Van Drunen, Sewalt, Oosterling, Weisbeek, Keultjes, Smeekens and Van Driel *et al.*, 1999). Using the FN50 criteria, the MATCH™ program correctly predicted all the experimentally confirmed S/MAR loci in the test sequences. The counts of flanking sequences containing S/MAR loci were 1119 and 1678 for chromosomes 4 and 5, respectively (Table 2). From this result, the densities of the S/MAR loci were calculated as one S/MAR locus per 15.5 kb for both chromosomes. This means two or three genes reside, on average, between two S/MAR loci, as the gene density is one per 4.6 kb and 4.4 kb for chromosomes 4 and 5, respectively (The European and The Coldspring Harbor 1999; The Kazusa DNA Research Institute 2002).

As a pivot, the S/MAR containing sequences give bigrams in both directions, so we collected the bigrams separately, before and after of the S/MAR containing flanking sequences. Additionally, we collected bigrams where the S/MAR resided in the middle of two genes. The proportions of functionally identical bigrams for chromosome 4 (Table 2) were higher (4.87~5.67%) than for chromosome 5 (2.06~2.41%). In chromosome 4, these

Table 2. Statistics on bigrams of chromosomes 4 and 5 considering S/MAR.

	Chromosome 4			Chromosome 5		
	Before S/MAR	Across S/MAR	After S/MAR	Before S/MAR	Across S/MAR	After S/MAR
No. of predicted S/MAR loci		1119		1678		
Expected No. of functionally identical bigrams	9.3483	8.7984	9.9899	9.6509	8.3922	8.9691
Observed No. of functionally identical bigrams	60 (4.87%)	65 (5.67%)	61 (4.87%)	43 (2.41%)	39 (2.28%)	37 (2.06%)
No. of bigrams w/o 00/98/99 categories	204 (16.55%)	192 (16.80%)	218 (17.41%)	184 (10.32%)	160 (9.37%)	171 (9.53%)
No of total Bigrams	1233	1143	1252	1783	1707	1794
P-value	$P(Z>16.9595) \approx 7.3690\times10^{-64}$	$P(Z>19.3969) \approx 3.8055\times10^{-83}$	$P(Z>16.5220) \approx 1.1369\times10^{-60}$	$P(Z>11.0280) \approx 1.0183\times10^{-27}$	$P(Z>10.8541) \approx 6.8648\times10^{-27}$	$P(Z>9.6153) \approx 2.2902\times10^{-21}$

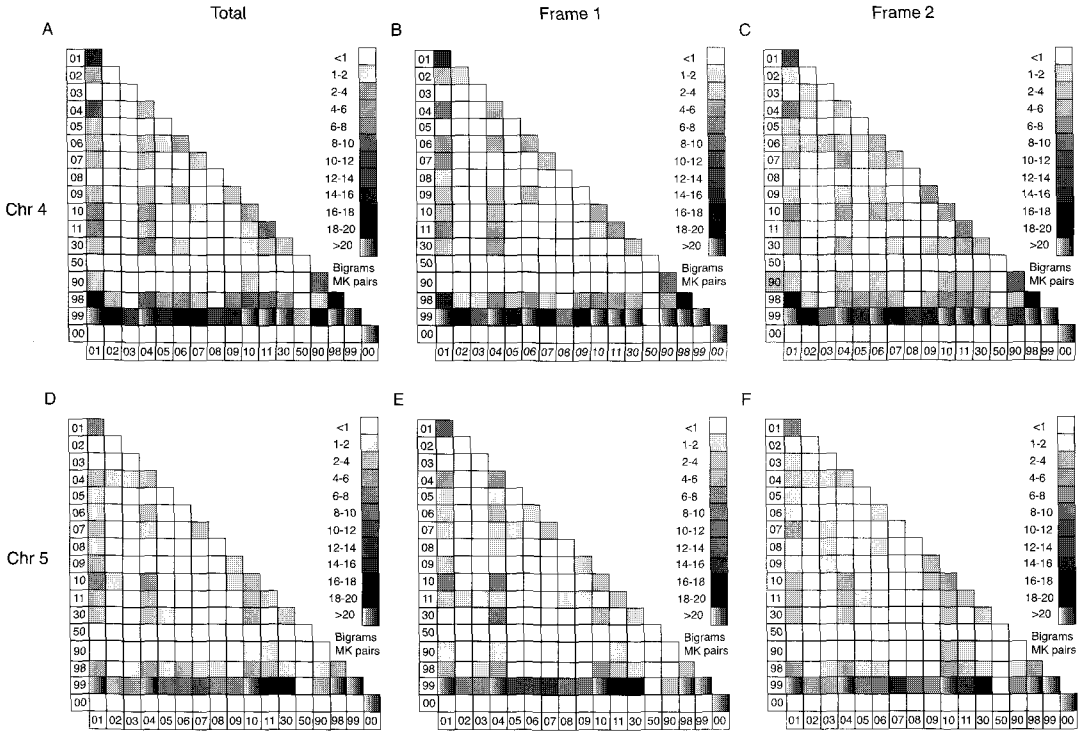

Fig. 2. The distributions of the functional combinations of neighboring genes when the S/MAR was not considered. Bigrams were mapped according to their nineteen large function categories. Panel A and C, indicated by 'Total', are the matrices from both bigram frames. Panel B and D are the matrices for the first bigram frame, and C and F are for the second bigram frame. The color gradient in the upper-right corner of each panel shows the bigram density per one thousand gene pairs. Numbers on the vertical and horizontal axis indicate the large functional categories.

proportions were similar, but slightly higher than in the case the S/MAR was not considered, especially in the class of 'Across S/MAR', suggesting S/MAR has some role in associating genes belonging to the same function category on the genome. In chromosome 5, however, the proportions were similar, but slightly lower than the cases that not consider the S/MAR, especially in the case of 'After S/MAR'. The p-values for functionally identical bigrams before, across and after the S/MAR on chromosome 4 (Table 2) were 7.3690×10^{-64}, 3.8055×10^{-83} and

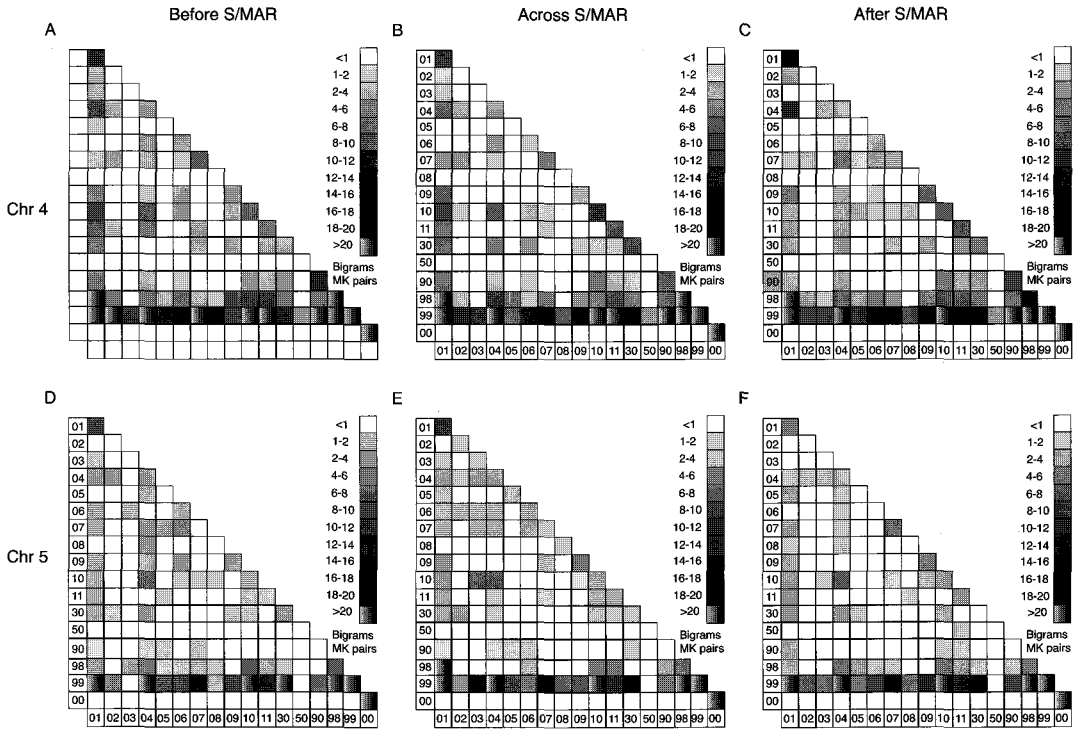

Fig. 3. The distributions of the functional combinations between neighboring genes when the S/MAR was considered. Bigrams were mapped according to their nineteen large function categories. Panel A and C are the matrices for bigrams located before S/MAR loci. Panel B and D are the matrices for bigrams with the S/MAR in the middle of them, and C and F are for the bigram located after the S/MAR. Numbers on the vertical and horizontal axis indicate the large functional categories.

Table 3. Statistics on trigrams of chromosome 4 and 5, without considering S/MAR.

	Frame 1	Frame 2	Frame 3	Total
Chromosome 4				
Expected No. of functionally identical trigrams	0.4540	0.4611	0.4824	
Observed No. of functionally identical trigrams	22 (1.46%)	24 (1.59%)	24 (1.60%)	70 (1.55%)
No. of trigrams w/o 00/98/99 categories	128 (8.49%)	130 (8.64%)	136 (9.05%)	394 (8.73%)
No. of total trigrams	1507	1505	1503	4515
P-value	$P(Z>32.0332) \approx 1.2700 \times 10^{-224}$	$P(Z>34.7262) \approx 8.4649 \times 10^{-264}$	$P(Z>33.9207) \approx 9.5589 \times 10^{-252}$	
Chromosome 5				
Expected No. of functionally identical trigrams	0.4611	0.4469	0.4256	
Observed No. of functionally identical trigrams	13 (0.59%)	15 (0.68%)	11 (0.50%)	39 (0.59%)
No. of trigrams w/o 00/98/99 categories	130 (5.91%)	126 (5.75%)	120 (5.47%)	376 (5.71%)
No. of total trigrams	2199	2190	2192	6581
P-value	$P(Z>18.4977) \approx 9.9222 \times 10^{-76}$	$P(Z>21.8087) \approx 8.9400 \times 10^{-105}$	$P(Z>16.2376) \approx 1.2125 \times 10^{-58}$	

Table 4. Statistics on trigrams considering S/MAR of chromosomes 4 and 5.

	Chromosome 4		Chromosome 5	
	Before S/MAR	After S/MAR	Before S/MAR	After S/MAR
Expected No. of functionally identical trigrams	0.4398	0.3937	0.3405	0.3476
Observed No. of functionally identical trigrams	23 (1.79%)	19 (1.48%)	12 (0.66%)	7 (0.38%)
No. of trigrams w/o 00/98/99 cat.	124 (9.66%)	111 (8.67%)	96 (5.26%)	98 (5.36%)
No. of total trigrams	1283	1281	1824	1828
P-value	$P(Z>34.0767) \approx 6.4265\times10^{-254}$	$P(Z>29.7065) \approx 2.4439\times10^{-193}$	$P(Z>20.0165) \approx 2.0943\times10^{-88}$	$P(Z>11.3029) \approx 3.0978\times10^{-29}$

1.1369×10^{-60}, respectively, and for chromosome 5 were 1.0183×10^{-27}, 6.8648×10^{-27} and 2.2902×10^{-21}, respectively. Although these *p*-values were much higher than those cases where the non-S/MAR were considered, it was difficult to determine if the S/MAR affects the genomic association of the bigrams, because *p*-values were all extremely low. Nevertheless, these *p*-values suggested there was little probability of the appearance due to chance in either case, and the observed number of functionally identical bigrams was significantly higher than expected considering the weighted random distribution. The matrices of these classes (Fig. 3) showed similar patterns to the cases where the S/MAR were not considered, and the diagonal pairs on chromosome 4 were denser than those on chromosome 5, as when the cases of the S/MAR was not considered.

Trigrams without considering S/MAR

As previously mentioned, the average interval of S/MAR loci in chromosomes 4 and 5 was 15.5 kb, and an average of two or three genes could reside in this interval. Thus, we extended the neighboring gene numbers to three, and assessed the association of three consecutive genes in their function. We divided cases into two classes, those where S/MAR were not considered and those where they were, for the analyses of trigrams.

For the cases where the S/MAR was not considered, we collected trigrams from three frames according to the start point. There were around 1500 trigrams for each frame in chromosome 4 and around 2190 trigrams in chromosome 5 (Table 3). The proportions of trigrams without the 00/98/99 categories, on average were 8.73 and 5.71% for chromosomes 4 and 5, respectively. This was about the half level of the bigrams because there were more chances of the 00, 98 and 99 function categories being neglected. The frequencies of functionally identical trigrams, on average were 1.55 and 0.59% for chromosomes 4 and 5. The *p*-values for the weighted random distributions were 1.2700×10^{-224}, 8.4649×10^{-264} and 9.5589×10^{-252} for frames 1, 2 and 3 of chromosome 4, and 9.9222×10^{-76}, 8.9400×10^{-105} and 1.2125×10^{-58} for chromosome 5, respectively. These *p*-values for both chromosomes suggested the probability of a genomic association of functionally identical trigram, due to chance, is extremely low, and the observed number of functionally identical trigrams was significantly higher, statistically, than expected assuming the same weighted random distribution as with the bigrams.

Trigrams considering S/MAR

Using the same predicted S/MAR loci information as for the analyses of the bigrams, we collected trigrams before and after S/MAR from each chromosome. The frequencies of the trigrams with identical function sub-category were 1.79 and 1.48% for the trigrams before and after the S/MAR position in chromosome 4, and 0.66 and 0.38% for chromosome 5 (Table 4), respectively, which were similar to those cases when S/MAR were not considered. The *p*-values were 6.4265×10^{-254} and 2.4439×10^{-193} before and after S/MAR on chromosome 4, and 2.0943×10^{-88} and 3.0978×10^{-29} on chromosome 5. This indicated the probability of a genomic association of functionally identical trigrams due to chance to be extremely low, and the observed number of functionally identical trigrams was significantly higher, statistically, than expected assuming the same weighted random distribution as for the bigrams. With these results, however, it was not possible to suggest any correlation between the genomic association of genes belonging to an identical function sub-category and S/MAR locus, because of the little difference in the frequency of functionally identical trigrams and *p*-values between cases when the S/MAR considered or not.

Discussion

The features of genes in eukaryotic genome are being revealed through the sequencing efforts, successive analyses by functional genomics and from *in silico* analysis. Operon-like structures of neighboring genes have been found in Caenorhabditis elegans (Blumental *et al.,* 2002), which suggests that similar organization could appear in the genome of other eukaryotic species. If those functionally related genes are assembled in a boundary on the genome, the regulation of their concerted expression at a higher level can be accomplished more easily, and the clustering of housekeeping genes of the human genome (Lercher *et al.,* 2002) can support this postulation. In addition, if there is any correlation between functions of the neighboring gene products, it could be used to predict both physical interactions between proteins, and protein function as the conservation of gene order in bacterial genomes is routinely used for the prediction of physical interactions of proteins, and the prediction of unknown function of neighboring gene (Dandekar *et al.,* 1998). However, investigation on this theme, have not been widely addressed on eukaryotic genomes.

In the present study, we described the association of genes belonging to identical function sub-categories on chromosomes 4 and 5 of *Arabidopsis thaliana*. We initiated this study by focusing on two consecutive gene sets, because the gene sets composed of two consecutive genes are the smallest of neighbored gene pairs, which we defined as 'bigrams' in this study. The collections of bigrams were divided into two cases according to the consideration of the S/MAR. The reason the S/MAR was considered for the collection of bigrams was as a result of the looped structure of interphase chromatin, caused by attachment of S/MAR on the nuclear matrix, can be dealt with a functional subunit, and therefore the S/MAR are thought to be the tools that subdivide eukaryotic genomes into structural and functional domains (Liebich *et al.,* 2002). Thus, before collecting the bigrams we predicted the S/MAR loci of the whole sequences on chromosomes 4 and 5 of *Arabidopsis thaliana*. The S/MARt DB, the database for S/MAR, contains information fully extracted from original publications. However, we could not find all the possible S/MAR loci of *Arabidopsis thaliana* from this database, due to the number of entries for plants only being 55, including 13 entries for *Arabidopsis thaliana*. Therefore, we predicted S/MAR loci from an *in silico* method. There are a couple of prediction tools publicly available such as MAR-Finder and SMARTest. They use several S/MAR related motifs in their predictions, but these features do not always appeared on every known S/MAR containing locus, and they are not adjusted to our subject, thus we had to devise another method. We used two MRS that had been reported as S/MAR motifs in *Arabidopsis thaliana* (Van Drunen *et al.,* 1997). These MRSs have been applied in other species (Van Drunen *et al.,* 1999), and furthermore, were defined as the most enriched motifs in S/MAR containing sequences (Liebich *et al.,* 2002). We extracted MRS matching sequences from 55 entries of dicotyledonous plants from the S/MARt DB, and made weight-matrices. These weight matrices were tested on several experimentally confirmed S/MAR containing sequences, and FN_50 profiles from the MATCH™ Profiler correctly predicted all of the S/MAR loci on them. The MATCH program predicted S/MAR loci on chromosomes 4 and 5, and the average interval between S/MAR loci was 15.5 kb for both chromosomes. With this length of sequences, an average of three genes can reside because the gene densities for chromosomes 4 and 5 are 4.6 kb and 4.4 kb per gene, respectively. Thus, we extended the range of analyses to three consecutive gene sets, which we defined as 'trigrams'.

In the collection where the S/MAR was not considered, we tried on different two frames for bigrams. In the first frame, we chose the first and second genes for the first bigram, and the third and fourth for the second, and so on. In the second frame, we chose the second and third genes for the first bigram and so on for subsequent bigrams. Similarly, we chose three frames for the trigrams. In this way, we collected independent bigrams and trigrams, and could calculate the p-values for the binomial random variable I, as described in materials and methods. The collections of bigrams and trigrams where the S/MAR was considered were also statistically independently extracted as they were separated by predefined S/MAR loci. Many of the collected bigrams and trigrams were excluded in this study, because we did not considered bigrams and trigrams with unknown or unclassified function categories. We calculated the proportions annotated genes, and the known function categories were assigned in the MAtDB, which were only about 32% (1190/3744) for chromosome 4 and 18% (1055/5874) for chromosome 5. The data for bigrams (Table 1 and 2) and trigrams (Table 3 and 4) showed the frequencies composing the genes belonged to identical function sub-categories, and were similar regardless of whether the S/MAR was considered or not, which was contrary to our expectations. The differences were only the frequencies of identical bigrams or trigrams of chromosome 5 were much smaller than chromosome 4. Although it was not easily possible to conclude, it might that there were more unknown or unclassified genes on chromosome 5. The p-values were evaluated to provide the statistical significance of the observed frequencies of functionally identical bigrams and trigrams.

We first calculated *p*-values for random uniform distributions, as with the report by Ge *et al.* (Ge *et al.*, 2001), where they evaluated *p*-values for the protein interacting pairs (and triplets) assuming each pair has the same probability in *Saccharomyces cerevisiae*. However, this assumption was not suitable for our study, because the *p*-values were extremely small when this assumption was made, and could not be used for calculations using our method. The other reason was that the proportion of annotated genes in the *Arabidopsis thaliana* genome are relatively small compared to those with the *Saccharomyces cerevisiae* genome whose gene functions are better understood, and furthermore their distribution is somewhat biased to a couple of function categories. Therefore, we provided another set of *p*-values for the weighted random distribution, and this assumption introduced a more realistic situation. In fact, excluding genes with unknown or unclassified functions, the function category for metabolism-01 was the largest in both chromosomes 4 and 5. We considered a set of genes with a known function sub-category, which we called *K*. Although the proportions of bigrams or trigrams consisting of an identical function sub-category were similar, of the total bigrams or trigrams available, when either the S/MAR was considered or not, the *p*-values were much different. If the S/MAR had some effect on the association of genes with respect to their function, the *p*-values when the S/MAR was considered should be much lower than when it is not, but the *p*-values when the S/MAR was considered were relatively higher. However, this did not mean the S/MAR affected negatively on the genomic association of genes with identical functions. This could be caused by differences in the number of bigrams or trigrams collected. If we were to try more bigrams (or trigrams), the situation becomes even further removed farther from the original assumption of the probability distribution - both for uniform and weighted random distribution. The real frequencies of the functionally identical bigrams and trigrams were much higher than expected, but the *p*-values suggested that these data were statistically significant. This suggested that regardless of the existence of S/MAR, there were significant associations of genes related in a certain cellular events on the genome. The clustering of housekeeping genes in human genomes has been reported, with suggestion that it might be advantageous to assemble housekeeping genes on some 'common ground' that remains in an open conformation across all cells (Lercher *et al.*, 2002). The analysis of co-expressed genes suggested the possibilities of grouping genes as a functional module (Thompson *et al.*, 2002), and the accumulation of such data will resolve the relation between the genomic association of genes and their functional significance. Additionally, more analyses on the link between the higher-order chromatin structure, and the gene clustering on the genome, should be addressed to prove this relationship.

Despite the inadequacy of the annotation information, this study has shown the significant association of neighboring genes with identical function sub-categories on chromosomes 4 and 5 of the *Arabidopsis thaliana* genome. Using all the information on genome annotation from large-scale functional genomics, the application of this strategy will reveal detailed and unbiased results, which complement experimental knowledge.

Materials and Methods

Data sources of *Arabidopsis thaliana* genome sequences and function annotation

Among five chromosomes of *Arabidopsis thaliana* we selected the chromosome 4 and 5 as the subject for analysis, as they are richer in annotation than the other three. The complete sequences of the two chromosomes were retrieved from the GenBank (Accession No. NC_003075.1 and NC_003076.2), and the function annotation information was retrieved from the Munich Information of Protein Sequences website (MIPS; http://mips.gsf.de/cgi-bin/proj/thal/search_funcat). From the GenBank flatfile, the features describing the sequence position and Arabidopsis Genome Initiative (AGI) code were extracted, and the AGI code linked to the annotation information and function category code from MIPS, such as enzyme category (EC) code. We used the function category codes, which *subdivided* 109 sub-categories from nineteen larger primary categories, including one additional customized category 'not found in the MAtDB,' which was assigned the code '00'. This information was saved in the form of a dictionary using an in-house program.

Collecting bigrams and trigrams

Pairs of two consecutive genes were collected from chromosomes 4 and 5 of *Arabidopsis thaliana*, and we refer to them as bigram in this study. Using similar strategy, three consecutive genes were collected, which we called trigrams. All the bigrams and trigrams were extracted from each frame according to their starting point. For example, bigrams from frame 1 were composed of the first-second, third-fourth, and so on, but those from frame 2 were composed of the second-third, fourth-fifth, and so on. We mapped *function category* code for the genes in the bigram and trigram using AGI code-function category code linking dictionary, then sorted them according to the number of function codes, and collected the ones not containing any '98: Classification not yet clear-cut', '99:

Unclassified proteins', or '00: Not found in MAtDB'. The bigrams and trigrams having an identical function sub-category were counted.

Prediction of scaffold/matrix attached region (S/MAR) loci

For the extraction of S/MAR motif weight matrices, we used S/MARt DB Professional 2.1 (Biobase GmbH, Germany; Release date: Jan. 21, 2002). We collected 55 entries corresponding to dicotyledonous plants, including 13 from *Arabidopsis thaliana*, from a total of 377 entries, and made them the subject for extracting the pattern of the 16-bp AWWRTAANNWWGNNNC and 8-bp AATAAYAA sequences, which were reported as the MAR/SAR recognition signature, (MRS)-1 and MRS-2, respectively (Van Drunen and Van Driel *et al.*, 1999). The MRS-1 and MRS-2 matching sequences were extracted using the MATCH™ program of TRANSFAC Professional 6.2 (Biobase GmbH, Germany), with a default motif core similarity of 75%. Their weight matrices were generated, using the MATCH™ Profiler program, from 79 MRS-1 and 19 MRS-2 matching sequences. The MRS were used as they give the advantage of increasing the chance of uncovering the S/MAR data that would otherwise be unavailable, and the enrichment of the MRS is higher than any other S/MAR *motifs* in experimentally confirmed sequences (Liebich and Wingender *et al.*, 2002). We selected the flanking sequences between non-overlapping genes as the targets for S/MAR prediction, but did not consider the S/MAR inside the coding sequences. These flanking sequences were extracted using positional information from the feature part of the GenBank flatfile. From these flanking sequences, both the MRS-1 and MRS-2 residing within 200 bp were colleted, without considering their orientation, using the MATCH™ program with a cutoff value of FN50 (MRS-1: 93%, MRS-2: 98.75%).

Collecting bigrams and trigrams considering S/MAR

For each predicted S/MAR, containing flanking sequences, the bigrams and trigrams at positions just before and after S/MAR were collected. In addition, bigrams that predicted where the S/MAR resides were also collected. We mapped the *function category* codes on these bigrams and trigrams, and then counted those that were functionally identical.

Statistical significance of bigrams and trigrams

To assess the statistical significance, *p-values* were calculated for the functionally identical bigrams. According to the Ge *et al.*, (2001) (Ge and Vidal *et al.*, 2001), *p*-values for protein-protein interactions in *Saccaromyces cerevisiae* were evaluated assuming each pair has the same probability - i.e. a uniform random distribution. We applied this concept with some modification, due to the bias in some of the distribution of function categories in the *Arabidopsis thaliana* genome. We calculated of *p*-values, excluding unknown function categories, but considered sets of genes with known function categories, which we called K.

The algorithm for the *p*-values of the weighted random distribution is as follows:

1. Estimate probabilities for each function sub-category of K.
   A. Count all the bigrams (or trigrams) for which both (or all) gene functions are known, N, and count the respective frequencies for each sub-category.
   B. Divide each frequency by the total number of available bigrams (or trigrams). The calculated results are the estimated probabilities.
2. Sum all the probabilities obtained to give the total probability of all the functionally identical bigrams (or trigrams), *p*.
3. Use a normal approximation to a binomial distribution to calculate *p*-values: Let I be the binomial random variable, with parameters *p* and i_0 being the true number of identical bigrams (or trigrams) in the data. The corresponding *p*-value is then given by the formula:

$$p = P(I > i_0) = \sum_{i=i_0+1}^{N} {}_N C_i \, p^i (1-p)^{N-i}$$

With respect to *p*, the expected number of functionally identical bigrams (or trigrams) is pN, and I is approximately normally distributed, $N(pN, p(1-p)N)$. Hence,

$$p \approx P\left(Z > \frac{i_0 - pN}{\sqrt{p(1-p)N}} \right)$$

where Z is a standard normal variable.

To show the distribution of the functional combination among neighboring genes, bigrams were assigned on the matrix according to their nineteen large function category codes. The frequencies of bigrams were normalized by scaling them down to 1000 pairs, and figured out as a twelve-color scale.

Acknowledgments

This research was supported by a grant from Plant Diversity Research Center of 21st Century Frontier Research Program funded by Ministry of Science and Technology of Korean government.

References

Blumental, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Theirry-Mieg, J., Chiu, W.L, Duke, K., Kiraly, M. and Kim, S. (2002) A global analysis of Caenorhabditis elegans operons. Nature, 417, 851 - 854.

Bode, J., Kohwi, Y., Dickinson, L., Joh, T., Klehr, D., Mielke, C. and Kohwi-Shigematsu, T. (1992) Biological significance of unwinding capability of nuclear matrix-associating DNAs. Science, 255, 195-197.

Cohen, B.A., Mitra, R.D., Hughes, J.D. and Church, G.M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nature Genet., 26, 183-186.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci., 23, 324-328.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature, 402, 86-90.

Frisch, M., Frech, K., Klingenhoff, A., Cartharius, K., Liebich, I. and Werner, T. (2001) In silico prediction of scaffold/matrix attachment regions in large genomic sequences. Genome Res., 12, 349-354.

Ge, H., Liu, Z., Church, G.M. and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nature Genet., 29, 482-486.

Huynen, M., Snel, B., Lathelll, W. and Bork, P. (2000) Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. Genome Res., 10, 1204-1210.

Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2002) Clustering of housekeeping-genes provides a unified model of gene order in the human genome. Nature Genet., 31, 180-183.

Liebich, I., Bode, J., Frisch, M. and Wingender, E. (2002) S/MARt DB: a database on scaffold/matrix attached regions. Nucleic Acids Res., 20, 372-274.

Liebich, I., Bode, J., Reuter, I. and Wingender, E. (2002) Evaluation of sequence motifs found in scaffold/matrix-attached regions (S/MARs). Nucleic Acids Res., 30, 3433-3442.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science, 285, 751-753.*

Mayer, k. et al., The European Union Arabidopsis Genome Sequencing Consortium & The Coldspring Harbor, Washington University in St Louis and PE Biosystems Arabidopsis Sequencing Consortium. (1999) Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. Nature, 402, 769-777.

Overbeek, R., Fonstein, M., D' Souza, M., Pusch, G.D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA, 96, 2896-2901.

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

Singh, G.B., Kramer, J.A. and Krawetz, S.A. (1997) Mathematical model to predict regions of chromatin attachment to the nuclear matrix. Nucleic Acids Res., 25, 1419-1425.

Snel, B., Bork, P. and Huynen, M.A. (2002) The identification of functional modules from the genomic association of genes. Proc. Natl. Acad. Sci. USA, 99, 5890-5895.

Stein, G.S. (1998) Interrelationships of nuclear architecture with gene expression: Functional encounters on a long and winding road. J. Cell. Biochem., 70, 157-158.

The Kazusa DNA Research Institute, The Cold Spring Harbor and Washington University in St Louis Sequencing Consortium and The European Union Arabidopsis Genome Sequencing Consortium. (2000) Sequence and analysis of chromosome 5 of the plant Arabidopsis thaliana. Nature, 408, 823-826.

Thompson, H.G.R., Harris, J.W., Wold, B.J., Quake, S.R. and Brody, J.P. (2002) Identification and confirmation of a module of coexpressed genes. Genome Res., 12, 1517-1522.

Van Drunen, C.M., Oosterling, R.W., Keultjes, G.M., Weisbeek, P.J, Van Driel, R. and Smeekens, S.C.M. (1997) Analysis of the chromatin domain organization around the platocyanin gene reveals an MAR-specific sequence element in Arabidopsis thaliana. Nucleic Acids Res., 25, 3904-3911.

Van Drunen, C.M., Sewalt, R.G.A.B., Oosterling, R.W., Weisbeek, P.J, Keultjes, G.M., Smeekens, S.C.M. and Van Driel, R. (1999) A bipartite sequence element associated with matrix/scaffold attachment regions. Nucleic Acids Res., 27, 2924-2930.

Von Mering, C. and Bork, P. (2002) Genome organization: Teamed up for transcription. Nature, 417, 797-798.

Yanai, I., Derti, A. and DeLisi, C. (2001) Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. Proc. Natl. Acad. Sci. USA, 98, 7940-7945.